

# The Global Optimization Geometry of Nonsymmetric Matrix Factorization and Sensing

Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B. Wakin\*

Department of Electrical Engineering and Computer Science  
Colorado School of Mines

March 6, 2017

## Abstract

In this paper we characterize the optimization geometry of a matrix factorization problem where we aim to find  $n \times r$  and  $m \times r$  matrices  $\mathbf{U}$  and  $\mathbf{V}$  such that  $\mathbf{UV}^T$  approximates a given matrix  $\mathbf{X}^*$ . We show that the objective function of the matrix factorization problem has no spurious local minima and obeys the strict saddle property not only for the exact-parameterization case where  $\text{rank}(\mathbf{X}^*) = r$ , but also for the over-parameterization case where  $\text{rank}(\mathbf{X}^*) < r$  and under-parameterization case where  $\text{rank}(\mathbf{X}^*) > r$ . These geometric properties imply that a number of iterative optimization algorithms (such as gradient descent) converge to a global solution with random initialization. For the exact-parameterization case, we further show that the objective function satisfies the robust strict saddle property, ensuring global convergence of many local search algorithms in polynomial time. We extend the geometric analysis to the matrix sensing problem with the factorization approach and prove that this global optimization geometry is preserved as long as the measurement operator satisfies the standard restricted isometry property.

## 1 Introduction

Low-rank matrices arise in a wide variety of applications throughout science and engineering, ranging from quantum tomography [1], signal processing [31], machine learning [35], and so on; see [16] for a comprehensive review. In all of these settings, we often encounter the following rank-constrained optimization problem:

$$\begin{aligned} & \underset{\mathbf{X} \in \mathbb{R}^{n \times m}}{\text{minimize}} \quad f(\mathbf{X}), \\ & \text{subject to} \quad \text{rank}(\mathbf{X}) \leq r, \end{aligned} \tag{1}$$

where the objective function  $f : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}$  is convex and smooth.

Despite the convexity of the objective function  $f$ , the rank constraint renders low-rank matrix optimizations of the form (1) highly nonconvex and computationally NP-hard in general [17]. Significant efforts have been devoted to transforming (1) into a convex problem by replacing the rank constraint with one involving the so-called nuclear norm. This strategy has been widely utilized in matrix inverse problems [34] arising in signal processing [16], machine learning [20], and control [17]. With convex analysis techniques, nuclear norm minimization has been proved to provide optimal performance in recovering low-rank matrices [11]. However, in spite of the optimal performance, solving nuclear norm minimization is very computationally expensive even with specialized first-order algorithms. For example, the singular value thresholding algorithm [7] requires performing an expensive singular value decomposition (SVD) in each iteration, making it computationally prohibitive in large-scale settings. This prevents nuclear norm minimization from scaling to practical problems.

---

\*Email: {zzhu, qiuli, gtang, mwakin}@mines.edu. This work was supported by NSF grant CCF-1409261, NSF grant CCF-1464205, and NSF CAREER grant CCF-1149225.

To relieve the computational bottleneck, recent studies propose to factorize the variable into  $\mathbf{X} = \mathbf{U}\mathbf{V}^T$ , and optimize over the  $n \times r$  and  $m \times r$  matrices  $\mathbf{U}$  and  $\mathbf{V}$  rather than the  $n \times m$  matrix  $\mathbf{X}$ . The rank constraint in (1) then is automatically satisfied through the factorization. This strategy is usually referred to as the Burer-Monteiro type decomposition after the authors in [5, 6]. Plugging this parameterization of  $\mathbf{X}$  in (1), we can recast the program into the following one:

$$\underset{\mathbf{U} \in \mathbb{R}^{n \times r}, \mathbf{V} \in \mathbb{R}^{m \times r}}{\text{minimize}} \quad h(\mathbf{U}, \mathbf{V}) := f(\mathbf{U}\mathbf{V}^T). \quad (2)$$

The bilinear nature of the parameterization renders the objective function of (2) nonconvex. Hence, it can potentially have spurious local minima (i.e., local minimizers that are not global minimizers) or even saddle points. With technical innovations in analyzing the landscape of nonconvex functions, several recent works have shown that the factored objective function  $h(\mathbf{U}, \mathbf{V})$  in matrix inverse problems has no spurious local minima [4, 19, 33].

## 1.1 Summary of results and outline

In this paper, we provide a comprehensive geometric analysis for the following non-square low-rank matrix factorization problem: given  $\mathbf{X}^* \in \mathbb{R}^{n \times m}$ ,

$$\underset{\mathbf{U} \in \mathbb{R}^{n \times r}, \mathbf{V} \in \mathbb{R}^{m \times r}}{\text{minimize}} \quad \left\| \mathbf{U}\mathbf{V}^T - \mathbf{X}^* \right\|_F^2, \quad (3)$$

which can be viewed as an important foundation of many popular matrix factorization problems such as the matrix sensing problem and matrix completion.

Our work actually rests on the recent works [18, 26, 32, 37] ensuring a number of iterative optimization methods (such as gradient descent) converge to a local minimum with random initialization provided the problem satisfies the so-called strict saddle property (see Definition 3 in Section 2). If the objective function further obeys the robust strict saddle property [18] (see Definition 4 in Section 2) or belongs to the class of so-called  $\mathcal{X}$  functions [37], the recent works [18, 37] show that many local search algorithms can converge to a local minimum in polynomial time. The implications of this line of work have had a tremendous impact for a number of nonconvex problems in applied mathematics, signal processing, and machine learning.

We begin this paper in Section 2 with the notions of strict saddle, strict saddle property, and robust strict saddle property. Considering that many invariant functions are not strongly convex (or even convex) in any neighborhood around a local minimum point, we then provide a revised robust strict saddle property requiring a regularity condition (see Definition 8 in Section 2) rather than strong convexity near the local minimum points (which is one of the requirements for the strict saddle property). Furthermore, by the same convergence analysis in [18] for problems satisfying the robust strict saddle property, the stochastic gradient descent algorithm is also guaranteed to converge to a local minimum point in polynomial time for problems satisfying the revised robust strict saddle property.

In Section 3, we show that the low-rank matrix factorization problem (3) (with an additional regularizer, see Section 3 for the details) has no spurious local minima and obeys the strict saddle property—that is the objective function in (3) has a directional negative curvature at all critical points but local minima—not only for the exact-parameterization case where  $\text{rank}(\mathbf{X}^*) = r$ , but also for the over-parameterization case where  $\text{rank}(\mathbf{X}^*) < r$  and the under-parameterization case where  $\text{rank}(\mathbf{X}^*) > r$ . The strict saddle property and lack of spurious local minima ensure that a number of local search algorithms applied to the matrix factorization problem (3) converge to global optima which correspond to the best rank- $r$  approximation to  $\mathbf{X}^*$ . Further, we completely analyze the low-rank matrix factorization problem (3) for the exact-parameterization case and show it obeys the revised robust strict saddle property.

In Section 4, we then extend the optimization geometry analysis for (3) to the matrix sensing problem. Provided the measurement operator satisfies the restricted isometry property (RIP) [34], we show the optimization geometry for the low-rank matrix factorization problem (3) is also preserved for the matrix sensing problem. In the case of Gaussian measurements, as guaranteed by this robust strict saddle property, a number of iterative optimizations can find the unknown matrix  $\mathbf{X}^*$  of rank  $r$  in polynomial time with high probability when the number of measurements exceeds a constant times  $(n + m)r^2$ . We conclude in Section 5 with a final discussion.

## 1.2 Relation to existing work

Unlike the objective functions of convex optimizations that have simple landscapes, such as where all local minimizers are global ones, the objective functions of general nonconvex programs have much more complicated landscapes. In recent years, by exploiting the underlying optimization geometry, a surge of progress has been made in providing theoretical justifications for matrix factorization problems such as (2) using a number of previously heuristic algorithms (such as alternating minimization, gradient descent, and the trust region method). Typical examples include phase retrieval [9, 13, 38], blind deconvolution [27, 29] dictionary learning [2, 36] and matrix sensing and completion [19, 23, 24, 39, 40].

These iterative algorithms can be sorted into two categories based on whether a good initialization is required. One set of algorithms consist of two steps: initialization and local refinement. Provided the function satisfies a regularity condition or similar properties, a good guess lying in the attraction basin of the global optimum can lead to global convergence of the following iterative step. We can obtain such initializations by spectral methods for phase retrieval [9] and low-rank matrix recovery problems [3, 39–41]. As we have mentioned, a regularity condition is also adopted in the revised robust strict saddle property.

Another category of works attempt to analyze the landscape of the objective functions in a larger space rather than the regions near the global optima. We can further separate these approaches into two types based on whether they involve the strict saddle property or the robust strict saddle property. The strict saddle property and lack of spurious local minima are proved for low-rank, positive semidefinite (PSD) matrix recovery [4] and completion [19], PSD matrix optimization problems with generic objective functions [28], low-rank non-square matrix estimation from linear observations [33], and low-rank non-square optimization problems with generic objective functions [42]. The strict saddle property along with the lack of spurious local minima ensures a number of iterative algorithms such as gradient descent [18] and the trust region method [15] converge to the global minimum with random initialization [18, 26, 36].

A few other works which are closely related to our work attempt to study the global geometry by characterizing the landscapes of the objective functions in the whole space rather than the regions near the global optima or all the critical points. As we discussed before, a number of local search algorithms are guaranteed to find a local optimum (which is also the global optimum if there are no spurious local minima) because of this robust strict saddle property. In [18], the authors proved that tensor decomposition problems satisfy this robust strict saddle property. Sun et al. [38] studied the global geometry of the phase retrieval problem. The very recent work in [30] analyzed the global geometry for PSD low-rank matrix factorization of the form (3) and the related matrix sensing problem when the rank is exactly parameterized (i.e.,  $r = \text{rank}(\mathbf{X}^*)$ ). We extend this line by considering the nonsymmetric low-rank matrix factorization and sensing problems.

Finally, we remark that our work is most closely related to the recent works in low-rank matrix factorization of the form (3) and its variants [4, 19, 30, 33, 39, 40]. As we discussed before, most of these works except [30] only characterize the geometry either near the global optima or all the critical points. Instead, we characterize the global geometry for the general (rather than PSD) low-rank matrix factorization and sensing. Furthermore, we show that the objective function in (3) obeys the strict saddle property and has spurious local minima not only for exact-parameterization ( $r = \text{rank}(\mathbf{X}^*)$ ), but also for over-parameterization ( $r > \text{rank}(\mathbf{X}^*)$ ) and under-parameterization ( $r < \text{rank}(\mathbf{X}^*)$ ). The under-parameterization implies that we can find the best rank- $r$  approximation to  $\mathbf{X}^*$  by many efficient iterative optimization algorithms such as gradient descent.

## 1.3 Notation

Before proceeding, we first briefly introduce some notation used throughout the paper. The symbols  $\mathbf{I}$  and  $\mathbf{0}$  respectively represent the identity and zero matrices with appropriate sizes. Also  $\mathbf{I}_n$  is used to denote the  $n \times n$  identity matrix. The set of  $r \times r$  orthonormal matrices is denoted by  $\mathcal{O}_r := \{\mathbf{R} \in \mathbb{R}^{r \times r} : \mathbf{R}^T \mathbf{R} = \mathbf{I}\}$ . For any natural number  $n$ , we let  $[n]$  or  $1 : n$  denote the set  $\{1, 2, \dots, n\}$ . We use  $|\Omega|$  denote the cardinality (i.e., the number of elements) of a set  $\Omega$ . MATLAB notations are adopted for matrix indexing; that is, for the  $n \times m$  matrix  $\mathbf{A}$ , its  $(i, j)$ -th element is denoted by  $\mathbf{A}[i, j]$ , its  $i$ -th row (or column) is denoted by  $\mathbf{A}[i, :]$  (or  $\mathbf{A}[:, i]$ ), and  $\mathbf{A}[\Omega_1, \Omega_2]$  refers to a  $|\Omega_1| \times |\Omega_2|$  submatrix obtained by taking the elements in rows  $\Omega_1$  of columns  $\Omega_2$  of matrix  $\mathbf{A}$ . Here  $\Omega_1 \subset [n]$  and  $\Omega_2 \subset [m]$ . We use  $a \gtrsim b$  (or  $a \lesssim b$ ) to represent that there is a constant so that  $a \geq \text{Const} \cdot b$  (or  $a \leq \text{Const} \cdot b$ ).

If a function  $h(\mathbf{U}, \mathbf{V})$  has two arguments,  $\mathbf{U} \in \mathbb{R}^{n \times r}$  and  $\mathbf{V} \in \mathbb{R}^{m \times r}$ , we occasionally use the notation

$h(\mathbf{W})$  when we put these two arguments into a new one as  $\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$ . For a scalar function  $f(\mathbf{Z})$  with a matrix variable  $\mathbf{Z} \in \mathbb{R}^{n \times m}$ , its gradient is an  $n \times m$  matrix whose  $(i, j)$ -th entry is  $[\nabla f(\mathbf{Z})][i, j] = \frac{\partial f(\mathbf{Z})}{\partial \mathbf{Z}[i, j]}$  for all  $i \in \{1, 2, \dots, n\}, j \in \{1, 2, \dots, m\}$ . The Hessian of  $f(\mathbf{Z})$  can be viewed as an  $nm \times nm$  matrix  $[\nabla^2 f(\mathbf{Z})][i, j] = \frac{\partial^2 f(\mathbf{Z})}{\partial \mathbf{Z}[i] \partial \mathbf{Z}[j]}$  for all  $i, j \in \{1, \dots, nm\}$ , where  $\mathbf{z}[i]$  is the  $i$ -th entry of the vectorization of  $\mathbf{Z}$ . An alternative way to represent the Hessian is by a bilinear form defined via  $[\nabla^2 f(\mathbf{Z})](\mathbf{A}, \mathbf{B}) = \sum_{i,j,k,l} \frac{\partial^2 f(\mathbf{Z})}{\partial \mathbf{Z}[i,j] \partial \mathbf{Z}[k,l]} \mathbf{A}[i, j] \mathbf{B}[k, l]$  for any  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times m}$ . These two notations will be used interchangeably whenever the specific form can be inferred from context.

## 2 Preliminaries

In this section, we provide a number of important definitions in optimization and group theory. To begin, suppose  $h(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$  is twice differentiable.

**Definition 1** (Critical points). *A point  $\mathbf{x}$  is a critical point of  $h(\mathbf{x})$  if  $\nabla h(\mathbf{x}) = \mathbf{0}$ .*

**Definition 2** (Strict saddles; or ridable saddles in [36]). *A critical point  $\mathbf{x}$  is a strict saddle if the Hessian matrix evaluated at this point has a strictly negative eigenvalue, i.e.,  $\lambda_{\min}(\nabla^2 h(\mathbf{x})) < 0$ .*

**Definition 3** (Strict saddle property [18]). *A twice differentiable function satisfies the strict saddle property if each critical point either corresponds to a local minimum or is a strict saddle.*

Intuitively, the strict saddle property requires a function to have a directional negative curvature at all of the critical points but local minima. This property allows a number of iterative algorithms such as noisy gradient descent [18] and the trust region method [15] to further decrease the function value at all the strict saddles and thus converge to a local minimum.

In [18], the authors proposed a noisy gradient descent algorithm for the optimization of functions satisfying the robust strict saddle property.

**Definition 4** (Robust strict saddle property [18]). *Given  $\alpha, \gamma, \epsilon, \delta$ , a twice differentiable  $h(\mathbf{x})$  satisfies the  $(\alpha, \gamma, \epsilon, \delta)$ -robust strict saddle property if for every point  $\mathbf{x}$  at least one of the following applies:*

1. *There exists a local minimum point  $\mathbf{x}^*$  such that  $\|\mathbf{x}^* - \mathbf{x}\| \leq \delta$ , and the function  $h(\mathbf{x}')$  restricted to a  $2\delta$  neighborhood of  $\mathbf{x}^*$  (i.e.,  $\|\mathbf{x}^* - \mathbf{x}'\| \leq 2\delta$ ) is  $\alpha$ -strongly convex;*
2.  $\lambda_{\min}(\nabla^2 h(\mathbf{x})) \leq -\gamma$ ;
3.  $\|\nabla h(\mathbf{x})\| \geq \epsilon$ .

In words, the above robust strict saddle property says that for any point whose gradient is small, then either the Hessian matrix evaluated at this point has a strictly negative eigenvalue, or it is close to a local minimum point. Thus the robust strict saddle property not only requires that the function obeys the strict saddle property, but also that it is well-behaved (i.e., strongly convex) near the local minima and has large gradient at the points far way to the critical points.

Intuitively, when the gradient is large, the function value will decrease in one step by gradient descent; when the point is close to a saddle point, the noise introduced in the noisy gradient descent could help the algorithm escape the saddle point and the function value will also decrease; when the point is close to a local minimum point, the algorithm then converges to a local minimum. Ge et al. [18] rigorously showed that the noisy gradient descent algorithm (see [18, Algorithm 1]) outputs a local minimum in a polynomial number of steps if the function  $h(\mathbf{x})$  satisfies the robust strict saddle property.

It is proved in [18] that tensor decomposition problems satisfy this robust strict saddle property. However, requiring the local strong convexity prohibits the potential extension of the analysis in [18] for the noisy gradient descent algorithm to many other problems, for which it is not possible to be strongly convex in any neighborhood around the local minimum points. Typical examples include the matrix factorization problems due to the rotational degrees of freedom for any critical point. This motivates us to weaken the local strong convexity assumption relying on the approach used by [9, 39] and to provide the following revised robust strict saddle property for such problems. To that end, we list some necessary definitions related to groups and invariance of a function under the group action.

**Definition 5** (Definition 7.1 [14]). *A (closed) binary operation,  $\circ$ , is a law of composition that produces an element of a set from two elements of the same set. More precisely, let  $\mathcal{G}$  be a set and  $a_1, a_2 \in \mathcal{G}$  be arbitrary elements. Then  $(a_1, a_2) \rightarrow a_1 \circ a_2 \in \mathcal{G}$ .*

**Definition 6** (Definition 7.2 [14]). A **group** is a set  $\mathcal{G}$  together with a (closed) binary operation  $\circ$  such that for any elements  $a, a_1, a_2, a_3 \in \mathcal{G}$  the following properties hold:

- *Associative property:*  $a_1 \circ (a_2 \circ a_3) = (a_1 \circ a_2) \circ a_3$ .
- *There exists an identity element*  $e \in \mathcal{G}$  *such that*  $e \circ a = a \circ e = a$ .
- *There is an element*  $a^{-1} \in \mathcal{G}$  *such that*  $a^{-1} \circ a = a \circ a^{-1} = e$ .

With this definition, it is common to denote a group just by  $\mathcal{G}$  without saying the binary operation  $\circ$  when it is clear from the context.

**Definition 7.** Given a function  $h(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$  and a group  $\mathcal{G}$  of operators on  $\mathbb{R}^n$ , we say  $h$  is invariant under the group action (or under an element  $a$  of the group) if

$$h(a(\mathbf{x})) = h(\mathbf{x})$$

for all  $\mathbf{x} \in \mathbb{R}^n$  and  $a \in \mathcal{G}$ .

Suppose the group action also preserves the energy of  $\mathbf{x}$ , i.e.,  $\|a(\mathbf{x})\| = \|\mathbf{x}\|$  for all  $a \in \mathcal{G}$ . Since for any  $\mathbf{x} \in \mathbb{R}^n$ ,  $h(a(\mathbf{x})) = h(\mathbf{x})$  for all  $a \in \mathcal{G}$ , it is straightforward to stratify the domain of  $h(\mathbf{x})$  into equivalent classes. The vectors in each of these equivalent classes differ by a group action. One implication is that when considering the distance of two points  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , it would be helpful to use the distance between their corresponding classes:

$$\text{dist}(\mathbf{x}_1, \mathbf{x}_2) := \min_{a_1 \in \mathcal{G}, a_2 \in \mathcal{G}} \|a_1(\mathbf{x}_1) - a_2(\mathbf{x}_2)\| = \min_{a \in \mathcal{G}} \|\mathbf{x}_1 - a(\mathbf{x}_2)\|, \quad (4)$$

where the second equality follows because  $\|a_1(\mathbf{x}_1) - a_2(\mathbf{x}_2)\| = \|a_1(\mathbf{x}_1 - a_1^{-1} \circ a_2(\mathbf{x}_2))\| = \|\mathbf{x}_1 - a_1^{-1} \circ a_2(\mathbf{x}_2)\|$  and  $a_1^{-1} \circ a_2 \in \mathcal{G}$ . Another implication is that the function  $h(\mathbf{x})$  cannot possibly be strongly convex (or even convex) in any neighborhood around its local minimum points because of the existence of the equivalent classes. Before presenting the revised robust strict saddle property for invariant functions, we list two examples to illuminate these concepts.

*Example 1:* As one example, consider the phase retrieval problem of recovering an  $n$ -dimensional complex vector  $\mathbf{x}^*$  from  $\{y_i = |\mathbf{b}_i^H \mathbf{x}^*|, i = 1, \dots, p\}$ , the magnitude of its projection onto a collection of known complex vectors  $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_p$  [9, 38]. The unknown  $\mathbf{x}^*$  can be estimated by solving the following natural least-squares formulation [9, 38]

$$\underset{\mathbf{x} \in \mathbb{C}^n}{\text{minimize}} \ h(\mathbf{x}) = \frac{1}{2p} \sum_{i=1}^p \left( y_i^2 - |\mathbf{b}_i^H \mathbf{x}|^2 \right)^2,$$

where we note that here the domain of  $\mathbf{x}$  is  $\mathbb{C}^n$ . For this case, we denote the corresponding

$$\mathcal{G} = \{e^{j\theta} : \theta \in [0, 1)\}$$

and the group action as  $a(\mathbf{x}) = e^{j\theta} \mathbf{x}$ , where  $a = e^{j\theta}$  is an element in  $\mathcal{G}$ . It is clear that  $h(a(\mathbf{x})) = h(\mathbf{x})$  for all  $a \in \mathcal{G}$ . Due to this invariance of  $h(\mathbf{x})$ , it is impossible to recover the global phase factor of the unknown  $\mathbf{x}^*$  and the function  $h(\mathbf{x})$  is not strongly convex in any neighborhood of  $\mathbf{x}^*$ .

*Example 2:* As another example, we revisit the general factored low-rank optimization problem (2):

$$\underset{\mathbf{U} \in \mathbb{R}^{n \times r}, \mathbf{V} \in \mathbb{R}^{m \times r}}{\text{minimize}} \ h(\mathbf{U}, \mathbf{V}) = f(\mathbf{U}\mathbf{V}^T).$$

We recast the two variables  $\mathbf{U}, \mathbf{V}$  into  $\mathbf{W}$  as  $\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$ . For this example, we denote the corresponding

$\mathcal{G} = \mathcal{O}_r$  and the group action on  $\mathbf{W}$  as  $a(\mathbf{W}) = \begin{bmatrix} \mathbf{U}\mathbf{R} \\ \mathbf{V}\mathbf{R} \end{bmatrix}$  where  $a = \mathbf{R} \in \mathcal{G}$ . We have that  $h(a(\mathbf{W})) = h(\mathbf{W})$

for all  $a \in \mathcal{G}$  since  $\mathbf{U}\mathbf{R}(\mathbf{V}\mathbf{R})^T = \mathbf{U}\mathbf{V}^T$  for any  $\mathbf{R} \in \mathcal{O}_r$ . Because of this invariance, in general  $h(\mathbf{W})$  is not strongly convex in any neighborhood around its local minimum points even though  $f(\mathbf{X})$  is a strongly convex function; see [30] for the symmetric low-rank factorization problem and Theorem 1 in Section 3 for the nonsymmetric low-rank factorization problem.

In the examples illustrated above, due to the invariance, the function is not strongly convex (or even convex) in any neighborhood around its local minimum point and thus it is prohibitive to apply the standard approach in optimization to show the convergence in a small neighborhood around the local minimum point. To overcome this issue, Candès et al. [9] utilized the so-called regularity condition as a sufficient condition for local convergence of gradient descent applied for the phase retrieval problem. This approach has also been applied for the matrix sensing problem [39] and semi-definite optimization [3].

**Definition 8** (Regularity condition [9, 39]). *Suppose  $h(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$  is invariant under the group action of the given group  $\mathcal{G}$ . Let  $\mathbf{x}^* \in \mathbb{R}^n$  be a local minimum point of  $h(\mathbf{x})$ . Define the set  $B(\delta, \mathbf{x}^*)$  as*

$$B(\delta, \mathbf{x}^*) := \{\mathbf{x} \in \mathbb{R}^n : \text{dist}(\mathbf{x}, \mathbf{x}^*) \leq \delta\},$$

where the distance  $\text{dist}(\mathbf{x}, \mathbf{x}^*)$  is defined in (4). Then we say the function  $h(\mathbf{x})$  satisfies the  $(\alpha, \beta, \delta)$ -regularity condition if for all  $\mathbf{x} \in B(\delta, \mathbf{x}^*)$ , we have

$$\langle \nabla h(\mathbf{x}), \mathbf{x} - a(\mathbf{x}^*) \rangle \geq \alpha \text{dist}(\mathbf{x}, \mathbf{x}^*)^2 + \beta \|\nabla h(\mathbf{x})\|^2, \quad (5)$$

where  $a = \arg \min_{a' \in \mathcal{G}} \|\mathbf{x} - a'(\mathbf{x}^*)\|$ .

We remark that  $(\alpha, \beta)$  in the regularity condition (8) must satisfy  $\alpha\beta \leq \frac{1}{4}$  since by applying Cauchy-Schwarz

$$\langle \nabla h(\mathbf{x}), \mathbf{x} - a(\mathbf{x}^*) \rangle \leq \|\nabla h(\mathbf{x})\| \text{dist}(\mathbf{x}, \mathbf{x}^*)$$

and the inequality of arithmetic and geometric means

$$\alpha \text{dist}^2(\mathbf{x}, \mathbf{x}^*) + \beta \|\nabla h(\mathbf{x})\|^2 \geq 2\sqrt{\alpha\beta} \text{dist}(\mathbf{x}, \mathbf{x}^*) \|\nabla h(\mathbf{x})\|^2.$$

**Lemma 1.** [9, 39] *If the function  $h(\mathbf{x})$  restricted to a  $\delta$  neighborhood of  $\mathbf{x}^*$  satisfies the  $(\alpha, \beta, \delta)$ -regularity condition, then as long as gradient descent starts from a point  $\mathbf{x}_0 \in B(\delta, \mathbf{x}^*)$ , the gradient descent update*

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \mu \nabla h(\mathbf{x}_t)$$

with step size  $0 < \nu \leq 2\beta$  obeys  $\mathbf{x}_t \in B(\delta, \mathbf{x}^*)$  and

$$\text{dist}^2(\mathbf{x}_t, \mathbf{x}^*) \leq (1 - 2\nu\alpha)^t \text{dist}^2(\mathbf{x}_0, \mathbf{x}^*)$$

for all  $t \geq 0$ .

The proof is given in [9]. To keep the paper self-contained, we also provide the proof of Lemma 1 in Appendix A. We remark that the decreasing rate  $1 - 2\nu\alpha \in [0, 1)$  since we choose  $\nu \leq 2\beta$  and  $\alpha\beta \leq \frac{1}{4}$ .

Now we establish the following revised robust strict saddle property for invariant functions by replacing the strong convexity condition in Definition 4 with the regularity condition.

**Definition 9** (Revised robust strict saddle property for invariant functions). *Given a twice differentiable  $h(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$  and a group  $\mathcal{G}$ , suppose  $h(\mathbf{x})$  is invariant under the group action and the energy of  $\mathbf{x}$  is also preserved under the group action, i.e.,  $h(a(\mathbf{x})) = h(\mathbf{x})$  and  $\|a(\mathbf{x})\|_2 = \|\mathbf{x}\|_2$  for all  $a \in \mathcal{G}$ . Given  $\alpha, \beta, \gamma, \epsilon, \delta$ ,  $h(\mathbf{x})$  satisfies the  $(\alpha, \beta, \gamma, \epsilon, \delta)$ -robust strict saddle property if for any point  $\mathbf{x}$  at least one of the following applies:*

1. *There exists a local minimum point  $\mathbf{x}^*$  such that  $\text{dist}(\mathbf{x}, \mathbf{x}^*) \leq \delta$ , and the function  $h(\mathbf{x}')$  restricted to  $2\delta$  a neighborhood of  $\mathbf{x}^*$  (i.e.,  $\text{dist}(\mathbf{x}', \mathbf{x}^*) \leq 2\delta$ ) satisfies the  $(\alpha, \beta, 2\delta)$ -regularity condition defined in Definition 8;*
2.  $\lambda_{\min}(\nabla^2 h(\mathbf{x})) \leq -\gamma$ ;
3.  $\|\nabla h(\mathbf{x})\| \geq \epsilon$ .

Compared with Definition 4, the revised robust strict saddle property requires the local descent condition instead of strict convexity in a small neighborhood around any local minimum point. With the convergence guarantee in Lemma 1, the convergence analysis of the stochastic gradient descent algorithm in [18] for the robust strict saddle functions can also be applied for the revised robust strict saddle functions defined in Definition 9 with the same convergence rate. We omit the details here. In the rest of the paper, the robust strict saddle property refers to the one in Definition 9.

### 3 The optimization geometry of low-rank matrix factorization

In this section, we consider the low-rank matrix factorization problem (3). Let  $\mathbf{X}^* = \Phi \Sigma \Psi^T = \sum_{i=1}^r \sigma_i \phi_i \psi_i^T$  be a reduced SVD of  $\mathbf{X}^*$ , where  $\Sigma$  is a diagonal matrix with  $\sigma_1 \geq \dots \geq \sigma_r$  along its diagonal. Denote  $\mathbf{U}^* = \Phi \Sigma^{1/2} \mathbf{R}$ ,  $\mathbf{V}^* = \Psi \Sigma^{1/2} \mathbf{R}$  for any  $\mathbf{R} \in \mathcal{O}_r$ . We first introduce the following ways to stack  $\mathbf{U}$  and  $\mathbf{V}$  together that are widely used through the paper:

$$\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}, \quad \widehat{\mathbf{W}} = \begin{bmatrix} \mathbf{U} \\ -\mathbf{V} \end{bmatrix}, \quad \mathbf{W}^* = \begin{bmatrix} \mathbf{U}^* \\ \mathbf{V}^* \end{bmatrix}, \quad \widehat{\mathbf{W}}^* = \begin{bmatrix} \mathbf{U}^* \\ -\mathbf{V}^* \end{bmatrix}.$$

Before moving on, we note that for any solution  $(\mathbf{U}, \mathbf{V})$  to (3),  $(\mathbf{U}\mathbf{R}_1, \mathbf{V}\mathbf{R}_2)$  is also a solution to (3) for any  $\mathbf{R}_1, \mathbf{R}_2 \in \mathbb{R}^{r \times r}$  such that  $\mathbf{U}\mathbf{R}_1\mathbf{R}_2^T\mathbf{V}^T = \mathbf{U}\mathbf{V}^T$ . As an extreme example,  $\mathbf{R}_1 = c\mathbf{I}$  and  $\mathbf{R}_2 = \frac{1}{c}\mathbf{I}$  where  $c$  can be arbitrarily large. In order to address this ambiguity (i.e., to reduce the search space of  $\mathbf{W}$  for (3)), we utilize the trick in [33, 39, 40, 42] by introducing a regularizer  $g$  and turn to solve the following problem

$$\underset{\mathbf{U} \in \mathbb{R}^{n \times r}, \mathbf{V} \in \mathbb{R}^{m \times r}}{\text{minimize}} \quad g(\mathbf{W}) := \frac{1}{2} \left\| \mathbf{U}\mathbf{V}^T - \mathbf{X}^* \right\|_F^2 + \rho(\mathbf{W}). \quad (6)$$

where

$$\rho(\mathbf{W}) := \frac{\mu}{4} \left\| \mathbf{U}^T \mathbf{U} - \mathbf{V}^T \mathbf{V} \right\|_F^2.$$

We remark that  $\mathbf{W}^*$  is still a global minimizer to the factored problem (6) since both the first term and  $\rho(\mathbf{W})$  achieve their global minimum at  $\mathbf{W}^*$ . The regularizer  $\rho(\mathbf{W})$  is applied to force the difference between the two Gram matrices of  $\mathbf{U}$  and  $\mathbf{V}$  as small as possible. The global minimum of  $\rho(\mathbf{W})$  is 0, which is achieved when  $\mathbf{U}$  and  $\mathbf{V}$  have the same Gram matrices, i.e., when  $\mathbf{W}$  belongs to

$$\mathcal{E} := \left\{ \mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} : \mathbf{U}^T \mathbf{U} - \mathbf{V}^T \mathbf{V} = \mathbf{0} \right\}. \quad (7)$$

Informally, we can view (6) as finding a point from  $\mathcal{E}$  that also minimizes the first term in (6). This is rigorously established in Lemma 3.

#### 3.1 Relationship to PSD low-rank matrix factorization

The following result to some degree characterizes the relationship between the nonsymmetric low-rank matrix factorization problem (6) and the following PSD low-rank matrix factorization problem [30]:

$$\underset{\mathbf{U} \in \mathbb{R}^{n \times r}}{\text{minimize}} \quad \left\| \mathbf{U}\mathbf{U}^T - \mathbf{M} \right\|_F^2 \quad (8)$$

where  $\mathbf{M} \in \mathbb{R}^{n \times n}$  is a rank- $r$  PSD matrix.

**Lemma 2.** Suppose  $g(\mathbf{W})$  is defined as in (6) with  $\mu > 0$ . Then we have

$$g(\mathbf{W}) \geq \min\left\{\frac{\mu}{4}, \frac{1}{8}\right\} \left\| \mathbf{W}\mathbf{W}^T - \mathbf{W}^*\mathbf{W}^{*T} \right\|_F^2.$$

In particular, if we choose  $\mu = \frac{1}{2}$ , then we have

$$g(\mathbf{W}) = \frac{1}{8} \left\| \mathbf{W}\mathbf{W}^T - \mathbf{W}^*\mathbf{W}^{*T} \right\|_F^2 + \frac{1}{4} \left\| \mathbf{U}^T \mathbf{U} - \mathbf{V}^T \mathbf{V} \right\|_F^2.$$

The proof of Lemma 2 is given in Appendix B. Informally, Lemma 2 indicates that minimizing  $g(\mathbf{W})$  also results in minimizing  $\left\| \mathbf{W}\mathbf{W}^T - \mathbf{W}^*\mathbf{W}^{*T} \right\|_F^2$  (which is the same form as the objective function in (8)) and hence the distance between  $\mathbf{W}$  and  $\mathbf{W}^*$  (though  $\mathbf{W}^*$  is unavailable a priori). The global geometry for the PSD low-rank matrix factorization problem (8) is recently analyzed by Li et al. in [30].



### 3.2 Characterization of critical points

We first provide the gradient and Hessian expression for  $g(\mathbf{W})$ . The gradient of  $g(\mathbf{W})$  is given by

$$\begin{aligned}\nabla_U g(\mathbf{U}, \mathbf{V}) &= (\mathbf{U}\mathbf{V}^T - \mathbf{X}^*)\mathbf{V} + \mu\mathbf{U}(\mathbf{U}^T\mathbf{U} - \mathbf{V}^T\mathbf{V}), \\ \nabla_V g(\mathbf{U}, \mathbf{V}) &= (\mathbf{U}\mathbf{V}^T - \mathbf{X}^*)^T\mathbf{U} - \mu\mathbf{V}(\mathbf{U}^T\mathbf{U} - \mathbf{V}^T\mathbf{V}),\end{aligned}$$

which can be rewritten as

$$\nabla g(\mathbf{W}) = \begin{bmatrix} (\mathbf{U}\mathbf{V}^T - \mathbf{X}^*)\mathbf{V} \\ (\mathbf{U}\mathbf{V}^T - \mathbf{X}^*)^T\mathbf{U} \end{bmatrix} + \mu\widehat{\mathbf{W}}\widehat{\mathbf{W}}^T\mathbf{W}.$$

Standard computations give the Hessian quadrature form  $[\nabla^2 g(\mathbf{W})](\Delta, \Delta)$  for any  $\Delta = \begin{bmatrix} \Delta_U \\ \Delta_V \end{bmatrix} \in \mathbb{R}^{(n+m) \times r}$  (where  $\Delta_U \in \mathbb{R}^{n \times r}$  and  $\Delta_V \in \mathbb{R}^{m \times r}$ ) as

$$[\nabla^2 g(\mathbf{W})](\Delta, \Delta) = \left\| \Delta_U \mathbf{V}^T + \mathbf{U} \Delta_V^T \right\|_F^2 + 2 \left\langle \mathbf{U} \mathbf{V}^T - \mathbf{X}^*, \Delta_U \Delta_V^T \right\rangle + [\nabla^2 \rho(\mathbf{W})](\Delta, \Delta) \quad (9)$$

where

$$[\nabla^2 \rho(\mathbf{W})](\Delta, \Delta) = \mu \left( \left\langle \widehat{\mathbf{W}}^T \mathbf{W}, \widehat{\Delta}^T \Delta \right\rangle + \left\langle \widehat{\mathbf{W}} \widehat{\Delta}^T, \Delta \mathbf{W}^T \right\rangle + \left\langle \widehat{\mathbf{W}} \widehat{\mathbf{W}}^T, \Delta \Delta^T \right\rangle \right). \quad (10)$$

The following result establishes that any critical point  $\mathbf{W}$  of  $g(\mathbf{W})$  belongs to  $\mathcal{E}$  (that is  $\mathbf{U}$  and  $\mathbf{V}$  are balanced factors of their product  $\mathbf{U}\mathbf{V}^T$ ) for any  $\mu > 0$ .

**Lemma 3.** *Suppose  $g(\mathbf{W})$  is defined as in (6) with  $\mu > 0$ . Then any critical point  $\mathbf{W}$  of  $g(\mathbf{W})$  belongs to  $\mathcal{E}$ , i.e.,*

$$\nabla g(\mathbf{W}) = \mathbf{0} \quad \Rightarrow \quad \mathbf{U}^T \mathbf{U} = \mathbf{V}^T \mathbf{V}. \quad (11)$$

The proof of Lemma 3 is given in Appendix C. By Lemma 3, we can simplify the equations for critical points as follows

$$\nabla_U \rho(\mathbf{U}, \mathbf{V}) = \mathbf{U} \mathbf{U}^T \mathbf{U} - \mathbf{X}^* \mathbf{V} = \mathbf{0}, \quad (12)$$

$$\nabla_V \rho(\mathbf{U}, \mathbf{V}) = \mathbf{V} \mathbf{V}^T \mathbf{V} - \mathbf{X}^{*T} \mathbf{U} = \mathbf{0}. \quad (13)$$

Now suppose  $\mathbf{W}$  is a critical point of  $g(\mathbf{W})$ . We can apply the Gram-Schmidt process to orthonormalize the columns of  $\mathbf{U}$  such that  $\tilde{\mathbf{U}} = \mathbf{U}\mathbf{R}$ , where  $\tilde{\mathbf{U}}$  is orthogonal and  $\mathbf{R} \in \mathcal{O}_r = \{\mathbf{R} \in \mathbb{R}^{r \times r}, \mathbf{R}^T \mathbf{R} = \mathbf{I}\}$ . Also let  $\tilde{\mathbf{V}} = \mathbf{V}\mathbf{R}$ . Since  $\mathbf{U}^T \mathbf{U} = \mathbf{V}^T \mathbf{V}$ , we have  $\tilde{\mathbf{U}}^T \tilde{\mathbf{U}} = \tilde{\mathbf{V}}^T \tilde{\mathbf{V}}$ . Thus  $\tilde{\mathbf{V}}$  is also orthogonal. Noting that  $\mathbf{U}\mathbf{V}^T = \tilde{\mathbf{U}}\tilde{\mathbf{V}}^T$ , we conclude that  $g(\mathbf{W}) = g(\tilde{\mathbf{W}})$  and  $\tilde{\mathbf{W}}$  is also a critical point of  $g(\mathbf{W})$  since  $\nabla g(\tilde{\mathbf{U}}) = \nabla g(\mathbf{U})\mathbf{R} = \mathbf{0}$  and  $\nabla g(\tilde{\mathbf{V}}) = \nabla g(\mathbf{V})\mathbf{R} = \mathbf{0}$ . Also for any  $\Delta \in \mathbb{R}^{(n+m) \times r}$ , we have  $[\nabla^2 g(\mathbf{W})](\Delta, \Delta) = [\nabla^2 g(\tilde{\mathbf{W}})](\Delta \mathbf{R}, \Delta \mathbf{R})$ , indicating that  $\mathbf{W}$  and  $\tilde{\mathbf{W}}$  have the same Hessian information. Thus, without loss of generality, we assume  $\mathbf{U}$  and  $\mathbf{V}$  are orthogonal. With this, we use  $\mathbf{u}_i$  and  $\mathbf{v}_i$  to denote the  $i$ -th columns of  $\mathbf{U}$  and  $\mathbf{V}$ , respectively. It follows from  $\nabla g(\mathbf{W}) = \mathbf{0}$  that

$$\begin{aligned}\|\mathbf{u}_i\|^2 \mathbf{u}_i &= \mathbf{X}^* \mathbf{v}_i, \\ \|\mathbf{v}_i\|^2 \mathbf{v}_i &= \mathbf{X}^{*T} \mathbf{u}_i,\end{aligned}$$

which indicates that

$$(\mathbf{u}_i, \mathbf{v}_i) \in \left\{ (\sqrt{\lambda_1} \mathbf{p}_1, \sqrt{\lambda_1} \mathbf{q}_1), \dots, (\sqrt{\lambda_r} \mathbf{p}_r, \sqrt{\lambda_r} \mathbf{q}_r), (\mathbf{0}, \mathbf{0}) \right\}.$$

Thus we identify all the critical points of  $g(\mathbf{W})$  in the following lemma, which is formally proved with an algebraic approach in Appendix D.

**Lemma 4.** *Let  $\mathbf{X}^* = \Phi \Sigma \Psi^T = \sum_{i=1}^r \sigma_i \phi_i \psi_i^T$  be a reduced SVD of  $\mathbf{X}^*$  and  $g(\mathbf{W})$  be defined as in (6) with  $\mu > 0$ . Any  $\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$  is a critical point of  $g(\mathbf{W})$  if and only if  $\mathbf{W} \in \mathcal{C}$  with*

$$\mathcal{C} := \left\{ \mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} : \mathbf{U} = \Phi \Lambda^{1/2} \mathbf{R}, \mathbf{V} = \Psi \Lambda^{1/2} \mathbf{R}, \mathbf{R} \in \mathcal{O}_r, \Lambda \text{ is diagonal}, \Lambda \geq \mathbf{0}, (\Sigma - \Lambda) \Sigma = \mathbf{0} \right\}. \quad (14)$$



Intuitively, (14) means that a critical point  $\mathbf{W}$  of  $g(\mathbf{W})$  is one such that  $\mathbf{U}\mathbf{V}^\top$  is a rank- $\ell$  approximation to  $\mathbf{X}^*$  with  $\ell \leq r$  and  $\mathbf{U}$  and  $\mathbf{V}$  are equal factors of this rank- $\ell$  approximation. Let  $\lambda_1, \lambda_2, \dots, \lambda_r$  denote the diagonals of  $\mathbf{\Lambda}$ . Unlike  $\mathbf{\Sigma}$ , we note that these diagonals  $\lambda_1, \lambda_2, \dots, \lambda_r$  are not necessarily placed in decreasing or increasing order. Actually, this equation  $(\mathbf{\Sigma} - \mathbf{\Lambda})\mathbf{\Sigma} = \mathbf{0}$  is equivalent to

$$\lambda_i \in \{\sigma_i, 0\}$$

for all  $i \in \{1, 2, \dots, r\}$ . Further, we introduce the set of optimal solutions:

$$\mathcal{X} := \left\{ \mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} : \mathbf{U} = \mathbf{\Phi}\mathbf{\Sigma}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi}\mathbf{\Sigma}^{1/2}\mathbf{R}, \mathbf{R} \in \mathcal{O}_r \right\}. \quad (15)$$

It is clear that the set  $\mathcal{X}$  containing all the optimal solutions, the set  $\mathcal{C}$  containing all the critical points and the set  $\mathcal{E}$  containing all the points with balanced factors have the nesting relationship:  $\mathcal{X} \subset \mathcal{C} \subset \mathcal{E}$ . Before moving to the next section, we provide one more result regarding  $\mathbf{W} \in \mathcal{E}$ . The proof of the following result is given in Appendix E.

**Lemma 5.** For any  $\mathbf{\Delta} = \begin{bmatrix} \mathbf{\Delta}_U \\ \mathbf{\Delta}_V \end{bmatrix} \in \mathbb{R}^{(n+m) \times r}$  and  $\mathbf{W} \in \mathcal{E}$  where  $\mathcal{E}$  is defined in (7), we have

$$\|\mathbf{\Delta}_U \mathbf{U}^\top\|_F^2 + \|\mathbf{\Delta}_V \mathbf{V}^\top\|_F^2 = \|\mathbf{\Delta}_U \mathbf{V}^\top\|_F^2 + \|\mathbf{\Delta}_V \mathbf{U}^\top\|_F^2, \quad (16)$$

and

$$\nabla^2 \rho(\mathbf{W}) \succeq \mathbf{0}. \quad (17)$$

### 3.3 Strict saddle property

Lemma 5 implies that the Hessian of  $\rho(\mathbf{W})$  evaluated at any critical point  $\mathbf{W}$  is PSD, i.e.,  $\nabla^2 \rho(\mathbf{W}) \succeq \mathbf{0}$  for all  $\mathbf{W} \in \mathcal{C}$ . Despite this fact, the following result establishes the strict saddle property for  $g(\mathbf{W})$ .

**Theorem 1.** Let  $g(\mathbf{W})$  be defined as in (6) with  $\mu > 0$  and  $\text{rank}(\mathbf{X}^*) = r$ . Let  $\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$  be any critical point satisfying  $\nabla g(\mathbf{W}) = \mathbf{0}$ , i.e.,  $\mathbf{W} \in \mathcal{C}$ . Any  $\mathbf{W} \in \mathcal{C} \setminus \mathcal{X}$  is a strict saddle of  $g(\mathbf{W})$  satisfying

$$\lambda_{\min}(\nabla^2 g(\mathbf{W})) \leq -\frac{1}{2} \left\| \mathbf{W}\mathbf{W}^\top - \mathbf{W}^* \mathbf{W}^{*\top} \right\| \leq -\sigma_r(\mathbf{X}^*). \quad (18)$$

Furthermore,  $g(\mathbf{W})$  is not strongly convex at any global minimum point  $\mathbf{W} \in \mathcal{X}$ .

The proof of Theorem 1 is given in Appendix F. Theorem 1 actually implies that  $g(\mathbf{W})$  has no spurious local minima (since all local minima belong to  $\mathcal{X}$ ) and obeys the strict saddle property. With the strict saddle property and lack of spurious local minima for  $g(\mathbf{W})$ , the recent result by Lee et al. [26] ensures that gradient descent converges to a global minimizer almost surely with random initialization.

### 3.4 Extension to over-parameterized case: $\text{rank}(\mathbf{X}^*) < r$

In this section, we briefly discuss the over-parameterized scenario where the low-rank matrix  $\mathbf{X}^*$  has rank smaller than  $r$ . Similar to Theorem 1, the following result shows that the strict saddle property also holds in this case.

**Theorem 2.** Let  $\mathbf{X}^* = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{\Psi}^\top = \sum_{i=1}^{r'} \sigma_i \phi_i \psi_i^\top$  be a reduced SVD of  $\mathbf{X}^*$  with  $r' \leq r$ , and let  $g(\mathbf{W})$  be defined as in (6) with  $\mu > 0$ . Any  $\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$  is a critical point of  $g(\mathbf{W})$  if and only if  $\mathbf{W} \in \mathcal{C}$  with

$$\mathcal{C} := \left\{ \mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} : \mathbf{U} = \mathbf{\Phi}\mathbf{\Lambda}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi}\mathbf{\Lambda}^{1/2}\mathbf{R}, \mathbf{R}\mathbf{R}^\top = \mathbf{I}_{r'}, \mathbf{\Lambda} \text{ is diagonal}, \mathbf{\Lambda} \geq \mathbf{0}, (\mathbf{\Sigma} - \mathbf{\Lambda})\mathbf{\Sigma} = \mathbf{0} \right\}$$

Further, all the local minima (which are also global) belong to the following set

$$\mathcal{X} = \left\{ \mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} : \mathbf{U} = \mathbf{\Phi}\mathbf{\Sigma}^{1/2}\mathbf{R}, \mathbf{V} = \mathbf{\Psi}\mathbf{\Sigma}^{1/2}\mathbf{R}, \mathbf{R}\mathbf{R}^\top = \mathbf{I}_{r'} \right\}$$

Finally, any  $\mathbf{W} \in \mathcal{C} \setminus \mathcal{X}$  is a strict saddle of  $g(\mathbf{W})$  satisfying

$$\lambda_{\min}(\nabla^2 g(\mathbf{W})) \leq -\frac{1}{2} \left\| \mathbf{W}\mathbf{W}^\top - \mathbf{W}^* \mathbf{W}^{*\top} \right\| \leq -\sigma_{r'}(\mathbf{X}^*).$$

The proof of Theorem 2 is given in Appendix G.

### 3.5 Extension to under-parameterized case: $\text{rank}(\mathbf{X}^*) > r$

We further discuss the under-parameterized case where  $\text{rank}(\mathbf{X}^*) > r$ . In this case, (3) is also known as the low-rank approximation problem as the product  $\mathbf{U}\mathbf{V}^T$  forms a rank- $r$  approximation to  $\mathbf{X}^*$ . Similar to Theorem 1, the following result shows that the strict saddle property also holds for  $g(\mathbf{W})$  in this scenario.

**Theorem 3.** Let  $\mathbf{X}^* = \Phi \Sigma \Psi^T = \sum_{i=1}^{r'} \sigma_i \phi_i \psi_i^T$  be a reduced SVD of  $\mathbf{X}^*$  with  $r' > r$  and  $\sigma_r(\mathbf{X}^*) > \sigma_{r+1}(\mathbf{X}^*)$ .<sup>1</sup> Also let  $g(\mathbf{W})$  be defined as in (6) with  $\mu > 0$ . Any  $\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$  is a critical point of  $g(\mathbf{W})$  if and only if  $\mathbf{W} \in \mathcal{C}$  with

$$\mathcal{C} := \left\{ \mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} : \mathbf{U} = \Phi[:, \Omega] \Lambda^{1/2} \mathbf{R}, \mathbf{V} = \Psi[:, \Omega] \Lambda^{1/2} \mathbf{R}, \right. \\ \left. \Lambda = \Sigma[\Omega, \Omega], \mathbf{R} \mathbf{R}^T = \mathbf{I}_\ell, \Omega \subset \{1, 2, \dots, r'\}, |\Omega| = \ell \leq r \right\}$$

where we recall that  $\Phi[:, \Omega]$  is a submatrix of  $\Phi$  obtained by keeping the columns indexed by  $\Omega$  and  $\Sigma[\Omega, \Omega]$  is an  $\ell \times \ell$  matrix obtained by taking the elements of  $\Sigma$  in rows and columns indexed by  $\Omega$ .

Further, all local minima belong to the following set

$$\mathcal{X} = \left\{ \mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} : \mathbf{U} = \Phi[:, 1:r] \Lambda^{1/2} \mathbf{R}, \mathbf{V} = \Psi[:, 1:r] \Lambda^{1/2} \mathbf{R}, \Lambda = \Sigma[1:r, 1:r], \mathbf{R} \in \mathcal{O}_r \right\}.$$

Finally, any  $\mathbf{W} \in \mathcal{C} \setminus \mathcal{X}$  is a strict saddle of  $g(\mathbf{W})$  satisfying

$$\lambda_{\min}(\nabla^2 g(\mathbf{W})) \leq -(\sigma_r(\mathbf{X}^*) - \sigma_{r+1}(\mathbf{X}^*)).$$

The proof of Theorem 3 is given in Appendix H. It follows from Eckart-Young-Mirsky theorem [22] that for any  $\mathbf{W} \in \mathcal{X}$ ,  $\mathbf{U}\mathbf{V}^T$  is the best rank- $r$  approximation to  $\mathbf{X}^*$ . Thus, this strict saddle property ensures that the local search algorithms applied to the factored problem (6) converge to global optimum which corresponds to the best rank- $r$  approximation to  $\mathbf{X}^*$ .

### 3.6 Robust strict saddle property

We now consider the revised robust strict saddle property defined in Definition 9 for the low-rank matrix factorization problem (6). As guaranteed by Theorem 1,  $g(\mathbf{W})$  satisfies the strict saddle property for any  $\mu > 0$ . However, too small a  $\mu$  would make analyzing the robust strict saddle property difficult. To see this, we denote

$$f(\mathbf{W}) = \frac{1}{2} \left\| \mathbf{U}\mathbf{V}^T - \mathbf{X}^* \right\|_F^2$$

for convenience. Thus we can rewrite  $g(\mathbf{W})$  as the sum of  $f(\mathbf{W})$  and  $\rho(\mathbf{W})$ . Note that for any  $\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} \in \mathcal{C}$  where  $\mathcal{C}$  is the set of critical points defined in (14),  $\widetilde{\mathbf{W}} = \begin{bmatrix} \mathbf{U}\mathbf{M} \\ \mathbf{V}\mathbf{M}^{-1} \end{bmatrix}$  is a critical point of  $f(\mathbf{W})$  for any invertible  $\mathbf{M} \in \mathbb{R}^{r \times r}$ . This further implies that the gradient at  $\widetilde{\mathbf{W}}$  reduces to

$$\nabla g(\widetilde{\mathbf{W}}) = \nabla \rho(\widetilde{\mathbf{W}}),$$

which could be very small if  $\mu$  is very small since  $\rho(\mathbf{W}) = \frac{\mu}{4} \left\| \mathbf{U}^T \mathbf{U} - \mathbf{V}^T \mathbf{V} \right\|_F^2$ . On the other hand,  $\widetilde{\mathbf{W}}$  could be far away from any point in  $\mathcal{X}$  for some  $\mathbf{M}$  that is not well-conditioned. Therefore, we choose a proper  $\mu$  controlling the importance of the regularization term such that for any  $\mathbf{W}$  that is not close to the critical points  $\mathcal{X}$ ,  $g(\mathbf{W})$  has large gradient. Motivated by Lemma 2, we choose  $\mu = \frac{1}{2}$ .

We note that Theorem 1 states that  $g(\mathbf{W})$  is not strongly convex at any at any global minimum point  $\mathbf{W} \in \mathcal{X}$  because of the invariance property of  $g(\mathbf{W})$ . To overcome this issue, we recall the discussions in Section 2 about the revised robust strict saddle property for the invariant functions. To that end,

<sup>1</sup>If  $\sigma_{r_1} = \dots = \sigma_r = \dots = \sigma_{r_2}$  with  $r_1 \leq r \leq r_2$ , then the optimal rank- $r$  approximation to  $\mathbf{X}^*$  is not unique. For this case, the optimal solution set  $\mathcal{X}$  for the factorized problem needs to be changed correspondingly, but the main arguments still hold.

we follow the notion of the distance between equivalent classes for invariant functions defined in (4) and define the distance between  $\mathbf{W}_1$  and  $\mathbf{W}_2$  as follows

$$\text{dist}(\mathbf{W}_1, \mathbf{W}_2) := \min_{\mathbf{R}_1 \in \mathcal{O}_r, \mathbf{R}_2 \in \mathcal{O}_r} \|\mathbf{W}_1 \mathbf{R}_1 - \mathbf{W}_2 \mathbf{R}_2\|_F = \min_{\mathbf{R} \in \mathcal{O}_r} \|\mathbf{W}_1 - \mathbf{W}_2 \mathbf{R}\|_F. \quad (19)$$

For convenience, we also denote the best rotation matrix  $\mathbf{R}$  so that  $\|\mathbf{W}_1 - \mathbf{W}_2 \mathbf{R}\|_F$  achieves its minimum by  $\mathbf{R}(\mathbf{W}_1, \mathbf{W}_2)$ , i.e.,

$$\mathbf{R}(\mathbf{W}_1, \mathbf{W}_2) := \arg \min_{\mathbf{R}' \in \mathcal{O}_r} \|\mathbf{W}_1 - \mathbf{W}_2 \mathbf{R}'\|_F, \quad (20)$$

which is also known as the orthogonal Procrustes problem [21]. The solution to the above minimization problem is characterized by the following lemma.

**Lemma 6.** [21] *Let  $\mathbf{W}_2^T \mathbf{W}_1 = \mathbf{L} \mathbf{S} \mathbf{P}^T$  be an SVD of  $\mathbf{W}_2^T \mathbf{W}_1$ . An optimal solution for the orthogonal Procrustes problem (20) is given by*

$$\mathbf{R}(\mathbf{W}_1, \mathbf{W}_2) = \mathbf{L} \mathbf{P}^T.$$

Moreover, we have

$$\mathbf{W}_1^T \mathbf{W}_2 \mathbf{R}(\mathbf{W}_1, \mathbf{W}_2) = (\mathbf{W}_2 \mathbf{R}(\mathbf{W}_1, \mathbf{W}_2))^T \mathbf{W}_1 = \mathbf{P} \mathbf{S} \mathbf{P}^T \succeq \mathbf{0}.$$

To ease the notation, we drop  $\mathbf{W}_1$  and  $\mathbf{W}_2$  in  $\mathbf{R}(\mathbf{W}_1, \mathbf{W}_2)$  and rewrite  $\mathbf{R}$  instead of  $\mathbf{R}(\mathbf{W}_1, \mathbf{W}_2)$  when they ( $\mathbf{W}_1$  and  $\mathbf{W}_2$ ) are clear from the context. Now we are well equipped to present the robust strict saddle property for  $g(\mathbf{W})$  in the following result.

**Theorem 4.** *Define the following regions*

$$\begin{aligned} \mathcal{R}_1 &:= \left\{ \mathbf{W} : \text{dist}(\mathbf{W}, \mathbf{W}^*) \leq \sigma_r^{1/2}(\mathbf{X}^*) \right\}, \\ \mathcal{R}_2 &:= \left\{ \mathbf{W} : \sigma_r(\mathbf{W}) \leq \sqrt{\frac{1}{2}} \sigma_r^{1/2}(\mathbf{X}^*), \|\mathbf{W} \mathbf{W}^T\|_F \leq \frac{20}{19} \|\mathbf{W}^* \mathbf{W}^{*T}\|_F \right\}, \\ \mathcal{R}_3' &:= \left\{ \mathbf{W} : \text{dist}(\mathbf{W}, \mathbf{W}^*) > \sigma_r^{1/2}(\mathbf{X}^*), \sigma_r(\mathbf{W}) > \sqrt{\frac{1}{2}} \sigma_r^{1/2}(\mathbf{X}^*), \right. \\ &\quad \left. \|\mathbf{W}\| \leq \frac{20}{19} \|\mathbf{W}^*\|, \|\mathbf{W} \mathbf{W}^T\|_F \leq \frac{20}{19} \|\mathbf{W}^* \mathbf{W}^{*T}\|_F \right\}, \\ \mathcal{R}_3'' &:= \left\{ \mathbf{W} : \|\mathbf{W}\| > \frac{20}{19} \|\mathbf{W}^*\| = \frac{20}{19} \sqrt{2} \|\mathbf{X}^*\|^{1/2}, \|\mathbf{W} \mathbf{W}^T\|_F \leq \frac{20}{19} \|\mathbf{W}^* \mathbf{W}^{*T}\|_F \right\}, \\ \mathcal{R}_3''' &:= \left\{ \mathbf{W} : \|\mathbf{W} \mathbf{W}^T\|_F > \frac{10}{9} \|\mathbf{W}^* \mathbf{W}^{*T}\|_F = \frac{20}{9} \|\mathbf{X}^*\|_F \right\}. \end{aligned}$$

Let  $g(\mathbf{W})$  be defined as in (6) with  $\mu = \frac{1}{2}$ . Then  $g(\mathbf{W})$  has the following robust strict saddle property:

1. For any  $\mathbf{W} \in \mathcal{R}_1$ ,  $g(\mathbf{W})$  satisfies local regularity condition:

$$\langle \nabla g(\mathbf{W}), \mathbf{W} - \mathbf{W}^* \mathbf{R} \rangle \geq \frac{1}{32} \sigma_r(\mathbf{X}^*) \text{dist}^2(\mathbf{W}, \mathbf{W}^*) + \frac{1}{48 \|\mathbf{X}^*\|} \|\nabla g(\mathbf{W})\|_F^2, \quad (21)$$

where  $\text{dist}(\mathbf{W}, \mathbf{W}^*)$  and  $\mathbf{R}$  are defined in (19) and (20), respectively.

2. For any  $\mathbf{W} \in \mathcal{R}_2$ ,  $g(\mathbf{W})$  has a directional negative curvature:

$$\lambda_{\min}(\nabla^2 g(\mathbf{W})) \leq -\frac{1}{4} \sigma_r(\mathbf{X}^*). \quad (22)$$

3. For any  $\mathbf{W} \in \mathcal{R}_3 = \mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3'''$ ,  $g(\mathbf{W})$  has large gradient descent:

$$\|\nabla g(\mathbf{W})\|_F \geq \frac{1}{10} \sigma_r^{3/2}(\mathbf{X}^*), \quad \forall \mathbf{W} \in \mathcal{R}_3'; \quad (23)$$

$$\|\nabla g(\mathbf{W})\|_F > \frac{39}{800} \|\mathbf{W}\|^3, \quad \forall \mathbf{W} \in \mathcal{R}_3''; \quad (24)$$

$$\|\nabla g(\mathbf{W})\|_F > \frac{1}{20} \|\mathbf{W} \mathbf{W}^T\|_F^{3/2}, \quad \forall \mathbf{W} \in \mathcal{R}_3'''. \quad (25)$$

The proof is given in Appendix I. We present several remarks better illustrating the above robust strict saddle property to conclude this section.

*Remark 1.* Both the right hand sides of (24) and (25) depend on  $\mathbf{W}$ . We can further obtain lower bounds for them by utilizing the fact that  $\mathbf{W}$  is large enough in both regions  $\mathcal{R}_3''$  and  $\mathcal{R}_3'''$ . Specifically, noting that  $\|\mathbf{W}\| > \frac{20}{19}\|\mathbf{W}^*\|$  for all  $\mathbf{W} \in \mathcal{R}_3''$ , we have

$$\|\nabla g(\mathbf{W})\|_F > \frac{39}{780}\sqrt{2}\|\mathbf{X}^*\|^{3/2}$$

for any  $\mathbf{W} \in \mathcal{R}_3''$ . Similarly, it follows from the fact  $\|\mathbf{W}\mathbf{W}^T\|_F > \frac{20}{9}\|\mathbf{X}^*\|_F$  for all  $\mathbf{W} \in \mathcal{R}_3'''$  that

$$\|\nabla g(\mathbf{W})\|_F > \frac{\sqrt{20}}{27}\|\mathbf{X}^*\|_F^{3/2}$$

for any  $\mathbf{W} \in \mathcal{R}_3'''$ .

*Remark 2.* Recall that all the strict saddles of  $g(\mathbf{W})$  are actually rank deficient (see Theorem 1). Thus the region  $\mathcal{R}_2$  attempts to characterize all the neighbors of the saddle saddles by including all rank deficient points. Actually, (22) holds not only for  $\mathbf{W} \in \mathcal{R}_2$ , but for all  $\mathbf{W}$  such that  $\sigma_r(\mathbf{W}) \leq \sqrt{\frac{1}{2}}\sigma_r^{1/2}(\mathbf{X}^*)$ . The reason we add another constraint controlling the term  $\|\mathbf{W}^*\mathbf{W}^{*T}\|_F$  is to ensure this negative curvature property in the region  $\mathcal{R}_2$  also holds for the matrix sensing problem discussed in next section. This is the same reason we add two more constraints  $\|\mathbf{W}\| \leq \frac{20}{19}\|\mathbf{W}^*\|_F$  and  $\|\mathbf{W}\mathbf{W}^T\|_F \leq \frac{10}{9}\|\mathbf{W}^*\mathbf{W}^{*T}\|_F$  for the region  $\mathcal{R}_3'$ .

*Remark 3.* Note that

$$\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3' \supseteq \left\{ \mathbf{W} : \|\mathbf{W}\| \leq \frac{20}{19}\|\mathbf{W}^*\|_F, \|\mathbf{W}\mathbf{W}^T\|_F \leq \frac{10}{9}\|\mathbf{W}^*\mathbf{W}^{*T}\|_F \right\},$$

which further implies

$$\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3' \cup \mathcal{R}_3'' \supseteq \left\{ \mathbf{W} : \|\mathbf{W}\mathbf{W}^T\|_F \leq \frac{10}{9}\|\mathbf{W}^*\mathbf{W}^{*T}\|_F \right\}.$$

Thus, we conclude that  $\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3 = \mathbb{R}^{(n+m) \times r}$ . Now the convergence analysis of the stochastic gradient descent algorithm in [18] for the robust strict saddle functions also holds for  $g(\mathbf{W})$ .

## 4 Matrix sensing with the factorization approach

In this section, we extend the previous geometric analysis to the matrix sensing problem

$$\underset{\mathbf{U} \in \mathbb{R}^{n \times r}, \mathbf{V} \in \mathbb{R}^{m \times r}}{\text{minimize}} \quad G(\mathbf{W}) := \frac{1}{2} \left\| \mathcal{A}(\mathbf{U}\mathbf{V}^T - \mathbf{X}^*) \right\|_2^2 + \rho(\mathbf{W}), \quad (26)$$

where  $\mathcal{A} : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^p$  is a known measurement operator and  $\rho(\mathbf{W})$  is the regularizer used in (6) and repeated here:

$$\rho(\mathbf{W}) = \frac{\mu}{4} \left\| \mathbf{U}^T \mathbf{U} - \mathbf{V}^T \mathbf{V} \right\|_F^2.$$

To give a sense that the geometry result in Theorem 4 for  $g(\mathbf{W})$  is also possibly preserved for  $G(\mathbf{W})$ , we suppose  $\mathcal{A}$  is a random variable that takes values in the set of linear maps from  $\mathbb{R}^{n \times m}$  to  $\mathbb{R}^p$  and is nearly isometrically distributed as for all  $\mathbf{X} \in \mathbb{R}^{n \times m}$

$$\mathbb{E} [\|\mathcal{A}(\mathbf{X})\|^2] = \|\mathbf{X}\|_F^2,$$

which is equivalent to

$$\mathbb{E} [\mathcal{A}^* \mathcal{A}] = \mathbf{I}, \quad (27)$$

where  $\mathcal{A}^* : \mathbb{R}^p \rightarrow \mathbb{R}^{n \times m}$  is the adjoint of the operator  $\mathcal{A}$ . The derivative of  $G(\mathbf{W})$  is given by

$$\nabla G(\mathbf{W}) = \begin{bmatrix} \mathcal{A}^* \mathcal{A}(\mathbf{U}\mathbf{V}^T - \mathbf{X}^*) \mathbf{V} \\ (\mathcal{A}^* \mathcal{A}(\mathbf{U}\mathbf{V}^T - \mathbf{X}^*))^T \mathbf{U} \end{bmatrix} + \mu \widehat{\mathbf{W}} \widehat{\mathbf{W}}^T \mathbf{W}. \quad (28)$$

For any  $\Delta = \begin{bmatrix} \Delta_U \\ \Delta_V \end{bmatrix} \in \mathbb{R}^{(n+m) \times r}$ , algebraic calculation gives the Hessian quadrature form  $[\nabla^2 G(\mathbf{W})](\Delta, \Delta)$  as

$$\begin{aligned} & [\nabla^2 g(\mathbf{W})](\Delta, \Delta) \\ &= \left\| \mathcal{A} \left( \Delta_U \mathbf{V}^T + U \Delta_V^T \right) \right\|_F^2 + 2 \left\langle \mathcal{A} \left( U \mathbf{V}^T - \mathbf{X}^* \right), \mathcal{A} \left( \Delta_U \Delta_V^T \right) \right\rangle + [\nabla^2 \rho(\mathbf{W})](\Delta, \Delta), \end{aligned} \quad (29)$$

where  $[\nabla^2 \rho(\mathbf{W})](\Delta, \Delta)$  is defined in (10). Using the near isometry property of  $\mathcal{A}$  as defined in (27), we have

$$\begin{aligned} \mathbb{E}(G(\mathbf{W})) &= g(\mathbf{W}), \\ \mathbb{E}(\nabla G(\mathbf{W})) &= \nabla g(\mathbf{W}), \\ \mathbb{E}(\nabla^2 G(\mathbf{W})) &= \nabla^2 g(\mathbf{W}). \end{aligned}$$

In words, the above results indicate that  $g(\mathbf{W})$ ,  $\nabla g(\mathbf{W})$  and  $\nabla^2 g(\mathbf{W})$  are respectively the unbiased estimators of the objective (26), the gradient (28), and the Hessian (29) of the matrix sensing problem. Thus, it is also expected that  $G(\mathbf{W})$ ,  $\nabla G(\mathbf{W})$ , and  $\nabla^2 G(\mathbf{W})$  are close to their counterparts for the matrix factorization problem when the map  $\mathcal{A}$  works similar to an identity map. The following matrix Restricted Isometry Property (RIP) serves as a way to link the low-rank matrix factorization problem (6) with the matrix sensing problem (26).

**Definition 10** (Restricted Isometry Property (RIP) [12, 34]). *The map  $\mathcal{A} : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^p$  satisfies the  $r$ -RIP with constant  $\delta_r$ ,<sup>2</sup> if*

$$(1 - \delta_r) \|\mathbf{X}\|_F^2 \leq \|\mathcal{A}(\mathbf{X})\|^2 \leq (1 + \delta_r) \|\mathbf{X}\|_F^2 \quad (30)$$

holds for any  $n \times m$  matrix  $\mathbf{X}$  with  $\text{rank}(\mathbf{X}) \leq r$ .

The following result establishes a similar geometry property to Theorem 4 when  $\mathcal{A}$  satisfies the RIP.

**Theorem 5.** *Let  $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3', \mathcal{R}_3'', \mathcal{R}_3'''$  be the regions as defined in Theorem 4. Let  $G(\mathbf{W})$  be defined as in (26) with  $\mu = \frac{1}{2}$  and  $\mathcal{A}$  satisfying the  $4r$ -RIP with*

$$\delta_{4r} \lesssim \frac{\sigma_r^{3/2}(\mathbf{X}^*)}{\|\mathbf{X}^*\|_F \|\mathbf{X}^*\|^{1/2}}. \quad (31)$$

Then  $G(\mathbf{W})$  has the following robust strict saddle property:

1. For any  $\mathbf{W} \in \mathcal{R}_1$ ,  $G(\mathbf{W})$  satisfies local regularity condition:

$$\langle \nabla G(\mathbf{W}), \mathbf{W} - \mathbf{W}^* \rangle \gtrsim \sigma_r(\mathbf{X}^*) \text{dist}^2(\mathbf{W}, \mathbf{W}^*) + \frac{1}{\|\mathbf{X}^*\|} \|\nabla G(\mathbf{W})\|_F^2, \quad (32)$$

where  $\text{dist}(\mathbf{W}, \mathbf{W}^*)$  and  $\mathbf{R}$  are defined in (19) and (20), respectively.

2. For any  $\mathbf{W} \in \mathcal{R}_2$ ,  $G(\mathbf{W})$  has a directional negative curvature, i.e.,

$$\lambda_{\min}(\nabla^2 G(\mathbf{W})) \lesssim -\sigma_r(\mathbf{X}^*). \quad (33)$$

3. For any  $\mathbf{W} \in \mathcal{R}_3 = \mathcal{R}_3' \cup \mathcal{R}_3'' \cup \mathcal{R}_3'''$ ,  $G(\mathbf{W})$  has large gradient descent:

$$\|\nabla G(\mathbf{W})\|_F \gtrsim \sigma_r^{3/2}(\mathbf{X}^*), \quad \forall \mathbf{W} \in \mathcal{R}_3'; \quad (34)$$

$$\|\nabla G(\mathbf{W})\|_F \gtrsim \|\mathbf{W}\|^3, \quad \forall \mathbf{W} \in \mathcal{R}_3''; \quad (35)$$

$$\|\nabla G(\mathbf{W})\|_F \gtrsim \left\| \mathbf{W} \mathbf{W}^T \right\|_F^{3/2}, \quad \forall \mathbf{W} \in \mathcal{R}_3'''. \quad (36)$$

The proof of this result is given in Appendix J. The main proof strategy is to utilize the RIP inequality about the measurement operator  $\mathcal{A}$  to control the deviation between the gradient (and the Hessian) of the matrix sensing problem and the counterpart of the matrix factorization problem so that the landscape of the matrix sensing problem has a similar geometry property. Several remarks follow.

<sup>2</sup>By abuse of notation, we adopt the conventional notation  $\delta_r$  for the RIP constant. The subscript  $r$  can be used to distinguish the RIP constant  $\delta_r$  from  $\delta$  which is used as a small constant in Section 2.

*Remark 4.* The constants involved in Theorem 5 can be found in Appendix J through the proof. Theorem 5 states that the objective function for the matrix sensing problem also satisfies the robust strict saddle property when (31) holds. The requirement for  $\delta_{4r}$  in (31) can be weakened to ensure the properties of  $g(\mathbf{W})$  are preserved for  $G(\mathbf{W})$  in some regions. For example, the local regularity condition (32) holds when

$$\delta_{4r} \leq \frac{1}{50}$$

which is independent of  $\mathbf{X}^*$ . Note that Tu et al. [39, Section 5.4, (5.15)] provided a similar regularity condition. However, the result there requires  $\delta_{6r} \leq \frac{1}{25}$  and  $\text{dist}(\mathbf{W}, \mathbf{W}^*) \leq \frac{1}{2\sqrt{2}}\sigma_r(\mathbf{X}^*)$  which defines a smaller region than  $\mathcal{R}_1$ . Based on this local regularity condition, Tu et al. [39] showed that gradient descent with a good initialization (which is close enough to  $\mathbf{W}^*$ ) converges to the unknown matrix  $\mathbf{W}^*$  (and hence  $\mathbf{X}^*$ ). With the analysis of the global geometric structure in  $G(\mathbf{W})$ , Theorem 5 ensures that many local search algorithms can converge to the global optimum with random initializations. In particular, stochastic gradient descent when applied to the matrix sensing problem (26) is guaranteed to find the unknown matrix  $\mathbf{X}^*$  in polynomial time.

*Remark 5.* A Gaussian  $\mathcal{A}$  will have the RIP with high probability when the number of measurements  $p$  is comparable to the number of degrees of freedom in an  $n \times m$  matrix with rank  $r$ . By Gaussian  $\mathcal{A}$  we mean the  $\ell$ -th element in  $\mathbf{y} = \mathcal{A}(\mathbf{X})$ ,  $y_\ell$ , is given by

$$y_\ell = \langle \mathbf{X}, \mathbf{A}_\ell \rangle = \sum_{i=1}^n \sum_{j=1}^m \mathbf{X}[i, j] \mathbf{A}_\ell[i, j],$$

where the entries of each  $n \times m$  matrix  $\mathbf{A}_\ell$  are independent and identically distributed normal random variables with zero mean and variance  $\frac{1}{p}$ . Specifically, a Gaussian  $\mathcal{A}$  satisfies (30) with high probability when [10, 16, 34]

$$p \gtrsim r(n+m) \frac{1}{\delta_r^2}.$$

Now utilizing the inequality  $\|\mathbf{X}^*\|_F \leq \sqrt{r}\|\mathbf{X}^*\|$  for (31), we conclude that in the case of Gaussian measurements, the robust strict saddle property is preserved for the matrix sensing problem with high probability when the number of measurements exceeds a constant times  $(n+m)r^2\kappa(\mathbf{X}^*)^3$  where  $\kappa(\mathbf{X}^*) = \frac{\sigma_1(\mathbf{X}^*)}{\sigma_r(\mathbf{X}^*)}$ . This further implies that, when applying the stochastic gradient descent algorithm to the matrix sensing problem (26) with Gaussian measurements, we are guaranteed to find the unknown matrix  $\mathbf{X}^*$  in polynomial time with high probability when

$$p \gtrsim (n+m)r^2\kappa(\mathbf{X}^*)^3. \quad (37)$$

When  $\mathbf{X}^*$  is an  $n \times n$  PSD matrix, Li et al. [30] showed that the corresponding matrix sensing problem with Gaussian measurements has similar global geometry to the low-rank PSD matrix factorization problem when the number of measurements

$$p \gtrsim nr^2 \frac{\sigma_1^4(\mathbf{X}^*)}{\sigma_r^2(\mathbf{X}^*)}. \quad (38)$$

Comparing (37) with (38), we find both results for the number of measurements needed depend similarly on the rank  $r$ , but slightly differently on the spectrum of  $\mathbf{X}^*$ . However, because Theorem 5 depends on the RIP of the map  $\mathcal{A}$ , our result can be applied to other matrix sensing problems whose measurement operator is not necessarily from a Gaussian measurement ensemble.

## 5 Conclusion

We have considered the low-rank matrix factorization problem—an important foundation of many popular matrix inverse problems such as matrix sensing and completion—using a factorization approach. Although the problem is nonconvex due to the bilinear nature of the variables, we showed that the objective function is well-behaved: it has a directional negative curvature in the region containing all the saddle points, obeys a regularity condition in any neighborhood around the local minima (here any local

minimum is global), and has a large gradient at points far away from the critical points. These geometric properties ensure that a number of iterative optimization algorithms converge to a global solution in polynomial time with an arbitrary initialization. When a matrix sensing problem is solved with the factorization approach, its objective function also obeys this robust strict saddle property as long as the measurement operator satisfies the RIP. It would be of interest to extend our geometric analysis to other matrix inverse problems such as matrix completion with the factorization approach.

## A Proof of Lemma 1

Denote  $a_{\mathbf{x}, \mathbf{x}^*} = \arg \min_{a' \in \mathcal{G}} \|\mathbf{x} - a'(\mathbf{x}^*)\|$ . Utilizing the definition of distance in (4), the regularity condition (5) and the assumption that  $\mu \leq 2\beta$ , we have

$$\begin{aligned} \text{dist}^2(\mathbf{x}_{t+1}, \mathbf{x}^*) &= \|\mathbf{x}_{t+1} - a_{\mathbf{x}_{t+1}, \mathbf{x}^*}(\mathbf{x}^*)\|^2 \\ &\leq \|\mathbf{x}_t - \nu \nabla h(\mathbf{x}_t) - a_{\mathbf{x}_t, \mathbf{x}^*}(\mathbf{x}^*)\|^2 \\ &= \|\mathbf{x}_t - a_{\mathbf{x}_t, \mathbf{x}^*}(\mathbf{x}^*)\|^2 + \nu^2 \|\nabla h(\mathbf{x}_t)\|^2 - 2\nu \langle \mathbf{x}_t - a_{\mathbf{x}_t, \mathbf{x}^*}(\mathbf{x}^*), \nabla h(\mathbf{x}_t) \rangle \\ &\leq (1 - 2\nu\alpha) \text{dist}^2(\mathbf{x}_t, \mathbf{x}^*) - \nu(2\beta - \nu) \|\nabla h(\mathbf{x}_t)\|^2 \\ &\leq (1 - 2\nu\alpha) \text{dist}^2(\mathbf{x}_t, \mathbf{x}^*) \end{aligned}$$

where the fourth line uses the regularity condition (5) and the last line holds because  $\nu \leq 2\beta$ . Thus we conclude  $\mathbf{x}_t \in B(\delta)$  for all  $t \in \mathbb{N}$  if  $\mathbf{x}_0 \in B(\delta)$  by noting that  $0 \leq 1 - 2\nu\alpha < 1$  since  $\alpha\beta \leq \frac{1}{4}$  and  $\nu \leq 2\beta$ .

## B Proof of Lemma 2

We first rewrite the objective function  $g(\mathbf{W})$ :

$$\begin{aligned} g(\mathbf{W}) &= \frac{1}{2} \|\mathbf{U}\mathbf{V}^T - \mathbf{U}^*\mathbf{V}^{*T}\|_F^2 + \frac{\mu}{4} \|\mathbf{U}^T\mathbf{U} - \mathbf{V}^T\mathbf{V}\|_F^2 \\ &\geq \min\{\mu, \frac{1}{2}\} \left( \|\mathbf{U}\mathbf{V}^T - \mathbf{U}^*\mathbf{V}^{*T}\|_F^2 + \frac{1}{4} \|\mathbf{U}^T\mathbf{U} - \mathbf{V}^T\mathbf{V}\|_F^2 \right) \\ &= \min\{\mu, \frac{1}{2}\} \left( \frac{1}{4} \|\mathbf{W}\mathbf{W}^T - \mathbf{W}^*\mathbf{W}^{*T}\|_F^2 + g'(\mathbf{W}) \right), \end{aligned}$$

where the second line attains the equality when  $\mu = \frac{1}{2}$ , and  $g'(\mathbf{W})$  in the last line is defined as

$$g'(\mathbf{W}) := \frac{1}{2} \|\mathbf{U}\mathbf{V}^T - \mathbf{U}^*\mathbf{V}^{*T}\|_F^2 - \frac{1}{4} \|\mathbf{U}\mathbf{U}^T - \mathbf{U}^*\mathbf{U}^{*T}\|_F^2 - \frac{1}{4} \|\mathbf{V}\mathbf{V}^T - \mathbf{V}^*\mathbf{V}^{*T}\|_F^2 + \frac{1}{4} \|\mathbf{U}^T\mathbf{U} - \mathbf{V}^T\mathbf{V}\|_F^2.$$

We further show  $g'(\mathbf{W})$  is always nonnegative:

$$\begin{aligned} g'(\mathbf{W}) &= \frac{1}{2} \|\mathbf{U}\mathbf{V}^T - \mathbf{U}^*\mathbf{V}^{*T}\|_F^2 - \frac{1}{4} \|\mathbf{U}\mathbf{U}^T - \mathbf{U}^*\mathbf{U}^{*T}\|_F^2 - \frac{1}{4} \|\mathbf{V}\mathbf{V}^T - \mathbf{V}^*\mathbf{V}^{*T}\|_F^2 + \frac{1}{4} \|\mathbf{U}^T\mathbf{U} - \mathbf{V}^T\mathbf{V}\|_F^2 \\ &= \frac{1}{2} \|\mathbf{U}\mathbf{V}^T - \mathbf{U}^*\mathbf{V}^{*T}\|_F^2 + \frac{1}{2} \|\mathbf{U}^T\mathbf{U}^*\|_F^2 + \frac{1}{2} \|\mathbf{V}^T\mathbf{V}^*\|_F^2 \\ &\quad - \frac{1}{2} \text{trace}(\mathbf{U}^T\mathbf{U}\mathbf{V}^T\mathbf{V}) - \frac{1}{4} \|\mathbf{U}^*\mathbf{U}^{*T}\|_F^2 - \frac{1}{4} \|\mathbf{V}^*\mathbf{V}^{*T}\|_F^2 \\ &= \frac{1}{2} \|\mathbf{U}^T\mathbf{U}^* - \mathbf{V}^T\mathbf{V}^*\|_F^2 + \frac{1}{2} \|\mathbf{U}^*\mathbf{V}^{*T}\|_F^2 - \frac{1}{4} \|\mathbf{U}^*\mathbf{U}^{*T}\|_F^2 - \frac{1}{4} \|\mathbf{V}^*\mathbf{V}^{*T}\|_F^2 \\ &= \frac{1}{2} \|\mathbf{U}^T\mathbf{U}^* - \mathbf{V}^T\mathbf{V}^*\|_F^2 \geq 0, \end{aligned}$$

where the last line follows because  $\mathbf{U}^{*T}\mathbf{U}^* = \mathbf{V}^{*T}\mathbf{V}^*$ . Thus, we have

$$g(\mathbf{W}) \geq \min\{\frac{\mu}{4}, \frac{1}{8}\} \|\mathbf{W}\mathbf{W}^T - \mathbf{W}^*\mathbf{W}^{*T}\|_F^2,$$

and

$$g(\mathbf{W}) = \frac{1}{8} \|\mathbf{W}\mathbf{W}^T - \mathbf{W}^*\mathbf{W}^{*T}\|_F^2 + \frac{1}{4} \|\mathbf{U}^T\mathbf{U}^* - \mathbf{V}^T\mathbf{V}^*\|_F^2$$

if  $\mu = \frac{1}{2}$ .



## C Proof of Lemma 3

Any critical point (see Definition 1)  $\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$  satisfies  $\nabla g(\mathbf{W}) = \mathbf{0}$ , i.e.,

$$\nabla_{\mathbf{U}} g(\mathbf{U}, \mathbf{V}) = (\mathbf{U}\mathbf{V}^T - \mathbf{X}^*)\mathbf{V} + \mu\mathbf{U}(\mathbf{U}^T\mathbf{U} - \mathbf{V}^T\mathbf{V}) = \mathbf{0}, \quad (39)$$

$$\nabla_{\mathbf{V}} g(\mathbf{U}, \mathbf{V}) = (\mathbf{U}\mathbf{V}^T - \mathbf{X}^*)^T\mathbf{U} - \mu\mathbf{V}(\mathbf{U}^T\mathbf{U} - \mathbf{V}^T\mathbf{V}) = \mathbf{0}. \quad (40)$$

By (40), we obtain

$$\mathbf{U}^T(\mathbf{U}\mathbf{V}^T - \mathbf{X}^*) = \mu(\mathbf{U}^T\mathbf{U} - \mathbf{V}^T\mathbf{V})\mathbf{V}^T.$$

Multiplying (39) by  $\mathbf{U}^T$  and plugging in the expression for  $\mathbf{U}^T(\mathbf{U}\mathbf{V}^T - \mathbf{X}^*)$  from the above equation gives

$$(\mathbf{U}^T\mathbf{U} - \mathbf{V}^T\mathbf{V})\mathbf{V}^T\mathbf{V} + \mathbf{U}^T\mathbf{U}(\mathbf{U}^T\mathbf{U} - \mathbf{V}^T\mathbf{V}) = \mathbf{0},$$

which further implies

$$\mathbf{U}^T\mathbf{U}\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V}\mathbf{V}^T\mathbf{V}.$$

Note that  $\mathbf{U}^T\mathbf{U}$  and  $\mathbf{V}^T\mathbf{V}$  are the principal square roots (i.e., PSD square roots) of  $\mathbf{U}^T\mathbf{U}\mathbf{U}^T\mathbf{U}$  and  $\mathbf{V}^T\mathbf{V}\mathbf{V}^T\mathbf{V}$ , respectively. Utilizing the result that a PSD matrix has a unique principal square root [25], we obtain

$$\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V}$$

for any critical point  $\mathbf{W}$ .

## D Proof of Lemma 4

We first repeat that  $\mathbf{X}^* = \Phi\Sigma\Psi^T$  is a reduced SVD of  $\mathbf{X}^*$ . We separate  $\mathbf{U}$  into two parts—the projections onto the column space of  $\Phi$  and its orthogonal complement—by denoting  $\mathbf{U} = \Phi\Lambda_1^{1/2}\mathbf{R}_1 + \mathbf{E}_1$  with  $\mathbf{R}_1 \in \mathcal{O}_r$ ,  $\mathbf{E}_1^T\Phi = \mathbf{0}$  and  $\Lambda_1$  being a  $r \times r$  diagonal matrix with non-negative elements along its diagonal. Similarly, denote  $\mathbf{V} = \Psi\Lambda_2^{1/2}\mathbf{R}_2 + \mathbf{E}_2$ , where  $\mathbf{R}_2 \in \mathcal{O}_r$ ,  $\mathbf{E}_2^T\Psi = \mathbf{0}$ ,  $\Lambda_2$  is a  $r \times r$  diagonal matrix with non-negative elements along its diagonal. Recall that any critical point  $\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$  satisfies

$$\nabla_{\mathbf{U}} \rho(\mathbf{U}, \mathbf{V}) = \mathbf{U}\mathbf{U}^T\mathbf{U} - \mathbf{X}^*\mathbf{V} = \mathbf{0},$$

$$\nabla_{\mathbf{V}} \rho(\mathbf{U}, \mathbf{V}) = \mathbf{V}\mathbf{V}^T\mathbf{V} - \mathbf{X}^{*T}\mathbf{U} = \mathbf{0}.$$

Plugging  $\mathbf{U} = \Phi\Lambda_1^{1/2}\mathbf{R}_1 + \mathbf{E}_1$  and  $\mathbf{V} = \Psi\Lambda_2^{1/2}\mathbf{R}_2 + \mathbf{E}_2$  into the above equations gives

$$\Phi\Lambda_1^{3/2}\mathbf{R}_1 + \Phi\Lambda_1^{1/2}\mathbf{R}_1\mathbf{E}_1^T\mathbf{E}_1 + \mathbf{E}_1\mathbf{R}_1^T\Lambda_1\mathbf{R}_1 + \mathbf{E}_1\mathbf{E}_1^T\mathbf{E}_1 - \Phi\Sigma\Lambda_2^{1/2}\mathbf{R}_2 = \mathbf{0}, \quad (41)$$

$$\Psi\Lambda_2^{3/2}\mathbf{R}_2 + \Psi\Lambda_2^{1/2}\mathbf{R}_2\mathbf{E}_2^T\mathbf{E}_2 + \mathbf{E}_2\mathbf{R}_2^T\Lambda_2\mathbf{R}_2 + \mathbf{E}_2\mathbf{E}_2^T\mathbf{E}_2 - \Psi\Sigma\Lambda_1^{1/2}\mathbf{R}_1 = \mathbf{0}. \quad (42)$$

Since  $\mathbf{E}_1$  is orthogonal to  $\Phi$ , (41) further implies that

$$\Phi\Lambda_1^{3/2}\mathbf{R}_1 + \Phi\Lambda_1^{1/2}\mathbf{R}_1\mathbf{E}_1^T\mathbf{E}_1 - \Phi\Sigma\Lambda_2^{1/2}\mathbf{R}_2 = \mathbf{0}, \quad (43)$$

$$\mathbf{E}_1\mathbf{R}_1^T\Lambda_1\mathbf{R}_1 + \mathbf{E}_1\mathbf{E}_1^T\mathbf{E}_1 = \mathbf{0}. \quad (44)$$

From (44), we have

$$\langle \mathbf{E}_1\mathbf{R}_1^T\Lambda_1\mathbf{R}_1 + \mathbf{E}_1\mathbf{E}_1^T\mathbf{E}_1, \mathbf{E}_1 \rangle = \langle \mathbf{R}_1^T\Lambda_1\mathbf{R}_1, \mathbf{E}_1^T\mathbf{E}_1 \rangle + \|\mathbf{E}_1\|_F^2 = 0,$$

which further implies  $\|\mathbf{E}_1\|_F^2 = 0$  by noting that  $\langle \mathbf{R}_1^T\Lambda_1\mathbf{R}_1, \mathbf{E}_1^T\mathbf{E}_1 \rangle \geq 0$  since it is the inner product between two PSD matrices. Thus  $\mathbf{E}_1 = \mathbf{0}$ . With a similar argument we also have  $\mathbf{E}_2 = \mathbf{0}$ .

With  $\mathbf{E}_1 = \mathbf{E}_2 = \mathbf{0}$ , (43) reduces to

$$\Phi \Lambda_1^{3/2} \mathbf{R}_1 - \Phi \Sigma \Lambda_2^{1/2} \mathbf{R}_2 = \mathbf{0}.$$

Since  $\Phi$  is orthogonal and  $\mathbf{R}_1 \in \mathcal{O}_r$ , the above equation implies that

$$\Lambda_1^{3/2} = \Sigma \Lambda_2^{1/2} \mathbf{R}_2 \mathbf{R}_1^T.$$

Let  $\Omega$  denote the set of locations of the non-zero diagonals in  $\Lambda_2$ , i.e.,  $\Lambda_2[i, i] > 0$  for all  $i \in \Omega$ . Then  $[\mathbf{R}_1^T]_\Omega = [\mathbf{R}_2^T]_\Omega$  since otherwise  $\Sigma \Lambda_2^{1/2} \mathbf{R}_2 \mathbf{R}_1^T$  is not a diagonal matrix anymore. Then we have

$$\Lambda_1^{3/2} = \Sigma \Lambda_2^{1/2} \quad (45)$$

implying that the set of the locations of non-zero diagonals in  $\Lambda_1$  is identical to  $\Omega$ . A similar argument applied to (42) gives

$$\Lambda_2^{3/2} = \Sigma \Lambda_1^{1/2}. \quad (46)$$

Noting that (45) implies  $\Lambda_1^{3/2}[i, i] = \Sigma[i, i] \Lambda_2^{1/2}[i, i]$  and (46) implies  $\Lambda_2^{3/2}[i, i] = \Sigma[i, i] \Lambda_1^{1/2}[i, i]$ , for all  $i \in \Omega$  we have  $\Lambda_1[i, i] = \Lambda_2[i, i] = \Sigma[i, i]$ . For  $i \notin \Omega$ , we have  $\Lambda_1[i, i] = \Lambda_2[i, i] = 0$ . Thus  $\Lambda_1 = \Lambda_2$ . For convenience, denote  $\Lambda = \Lambda_1 = \Lambda_2$  with  $\Lambda[i, i] = \lambda_i$ .

Finally, we note that  $\mathbf{U} = \Phi \Lambda^{1/2} \mathbf{R}_1 = \sum_{i \in \Omega} \lambda_i \phi_i \mathbf{R}_1[i, :]$  and  $\mathbf{V} = \Psi \Lambda^{1/2} \mathbf{R}_2 = \sum_{i \in \Omega} \lambda_i \psi_i \mathbf{R}_2[i, :]$  implying that only  $[\mathbf{R}_1^T]_\Omega$  and  $[\mathbf{R}_2^T]_\Omega$  play a role in  $\mathbf{U}$  and  $\mathbf{V}$ , respectively. Thus one can set  $\mathbf{R}_1 = \mathbf{R}_2$  since we already proved  $[\mathbf{R}_1^T]_\Omega = [\mathbf{R}_2^T]_\Omega$ .

## E Proof of Lemma 5

Utilizing the result that any point  $\mathbf{W} \in \mathcal{E}$  satisfies  $\widehat{\mathbf{W}}^T \mathbf{W} = \mathbf{U}^T \mathbf{U} - \mathbf{V}^T \mathbf{V} = \mathbf{0}$ , we directly obtain

$$\|\Delta_U \mathbf{U}^T\|_F^2 + \|\Delta_V \mathbf{V}^T\|_F^2 = \|\Delta_U \mathbf{V}^T\|_F^2 + \|\Delta_V \mathbf{U}^T\|_F^2$$

since  $\|\Delta_U \mathbf{U}^T\|_F^2 = \text{trace}(\Delta_U \mathbf{U}^T \mathbf{U} \Delta_U) = \text{trace}(\Delta_U \mathbf{V}^T \mathbf{V} \Delta_U) = \|\Delta_U \mathbf{V}^T\|_F^2$  (and similarly for the other two terms).

We then rewrite the last two terms in (10) as

$$\begin{aligned} & \langle \widehat{\mathbf{W}} \widehat{\Delta}^T, \Delta \mathbf{W}^T \rangle + \langle \widehat{\mathbf{W}} \widehat{\mathbf{W}}^T, \Delta \Delta^T \rangle \\ &= \langle \widehat{\mathbf{W}}^T \Delta, \Delta^T \widehat{\mathbf{W}} \rangle + \langle \widehat{\mathbf{W}}^T \Delta, \widehat{\mathbf{W}}^T \Delta \rangle \\ &= \langle \widehat{\mathbf{W}}^T \Delta, \widehat{\mathbf{W}}^T \Delta + \Delta^T \widehat{\mathbf{W}} \rangle \\ &= \frac{1}{2} \langle \widehat{\mathbf{W}}^T \Delta + \Delta^T \widehat{\mathbf{W}}, \widehat{\mathbf{W}}^T \Delta + \Delta^T \widehat{\mathbf{W}} \rangle + \frac{1}{2} \langle \widehat{\mathbf{W}}^T \Delta - \Delta^T \widehat{\mathbf{W}}, \widehat{\mathbf{W}}^T \Delta + \Delta^T \widehat{\mathbf{W}} \rangle \\ &= \frac{1}{2} \left\| \widehat{\mathbf{W}}^T \Delta + \Delta^T \widehat{\mathbf{W}} \right\|_F^2 \end{aligned}$$

where the last line holds because  $\langle \mathbf{A} - \mathbf{A}^T, \mathbf{A} + \mathbf{A}^T \rangle = 0$ . Plugging these with the factor  $\widehat{\mathbf{W}}^T \mathbf{W} = \mathbf{0}$  into the Hessian quadrature form  $[\nabla^2 \rho(\mathbf{W})](\Delta, \Delta)$  defined in (10) gives

$$[\nabla^2 \rho(\mathbf{W})](\Delta, \Delta) \geq \frac{\mu}{2} \left\| \widehat{\mathbf{W}}^T \Delta + \Delta^T \widehat{\mathbf{W}} \right\|_F^2 \geq 0.$$

This implies that the Hessian of  $\rho$  evaluated at any  $\mathbf{W} \in \mathcal{E}$  is PSD, i.e.,  $\nabla^2 \rho(\mathbf{W}) \succeq \mathbf{0}$ .<sup>3</sup>

---

<sup>3</sup>This can also be observed since any critical point  $\mathbf{W}$  is a global minimum of  $\rho(\mathbf{W})$ , which directly indicates that  $\nabla^2 \rho(\mathbf{W}) \succeq \mathbf{0}$ .

## F Proof of Theorem 1 (strict saddle property for (6))

We begin the proof of Theorem 1 by characterizing any  $\mathbf{W} \in \mathcal{C} \setminus \mathcal{X}$ . For this purpose, let  $\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$ , where  $\mathbf{U} = \Phi \Lambda^{1/2} \mathbf{R}$ ,  $\mathbf{V} = \Psi \Lambda^{1/2} \mathbf{R}$ ,  $\mathbf{R} \in \mathcal{O}_r$ ,  $\Lambda$  is diagonal,  $\Lambda \geq \mathbf{0}$ ,  $(\Sigma - \Lambda)\Sigma = \mathbf{0}$ , and  $\text{rank}(\Lambda) < r$ . Denote the corresponding optimal solution  $\mathbf{W}^* = \begin{bmatrix} \mathbf{U}^* \\ \mathbf{V}^* \end{bmatrix}$ , where  $\mathbf{U}^* = \Phi \Sigma^{1/2} \mathbf{R}$ ,  $\mathbf{V}^* = \Psi \Sigma^{1/2} \mathbf{R}$ . Let

$$k = \arg \max_i \sigma_i - \lambda_i$$

denote the location of the first zero diagonal element in  $\Lambda$ . Noting that  $\lambda_i \in \{\sigma_i, 0\}$ , we conclude that

$$\lambda_k = 0, \quad \phi_k^T \mathbf{U} = \mathbf{0}, \quad \psi_k^T \mathbf{V} = \mathbf{0}. \quad (47)$$

In words,  $\phi_k$  and  $\psi_k$  are orthogonal to  $\mathbf{U}$  and  $\mathbf{V}$ , respectively. Let  $\alpha \in \mathbb{R}^r$  be the eigenvector associated with the smallest eigenvalue of  $\mathbf{W}^T \mathbf{W}$ . Such  $\alpha$  simultaneously lives in the null spaces of  $\mathbf{U}$  and  $\mathbf{V}$  since  $\mathbf{W}$  is rank deficient indicating

$$0 = \alpha^T \mathbf{W}^T \mathbf{W} \alpha = \alpha^T \mathbf{U}^T \mathbf{U} \alpha + \alpha^T \mathbf{V}^T \mathbf{V} \alpha,$$

which further implies

$$\begin{cases} \alpha^T \mathbf{U}^T \mathbf{U} \alpha = 0, \\ \alpha^T \mathbf{V}^T \mathbf{V} \alpha = 0. \end{cases} \quad (48)$$

With this property, we construct  $\Delta$  by setting  $\Delta_U = \phi_k \alpha^T$  and  $\Delta_V = \psi_k \alpha^T$ . Now we show that  $\mathbf{W}$  is a strict saddle by arguing that  $g(\mathbf{W})$  has a strictly negative curvature along the constructed direction  $\Delta$ , i.e.,  $[\nabla^2 g(\mathbf{W})](\Delta, \Delta) < 0$ . To that end, we compute the five terms in (9) as follows

$$\begin{aligned} \|\Delta_U \mathbf{V}^T + \mathbf{U} \Delta_V^T\|_F^2 &= 0 \quad (\text{since } (48)), \\ \langle \mathbf{U} \mathbf{V}^T - \mathbf{X}^*, \Delta_U \Delta_V^T \rangle &= \lambda_k - \sigma_k = -\sigma_k \quad (\text{since } (47)), \\ \langle \widehat{\mathbf{W}}^T \mathbf{W}, \widehat{\Delta}^T \Delta \rangle &= 0 \quad (\text{since } \widehat{\mathbf{W}}^T \mathbf{W} = \mathbf{0}), \\ \langle \widehat{\mathbf{W}} \widehat{\Delta}^T, \Delta \mathbf{W}^T \rangle &= \text{trace}(\widehat{\Delta}^T \mathbf{W} \Delta^T \widehat{\mathbf{W}}) = 0 \quad (\text{since } \widehat{\Delta}^T \mathbf{W} = \mathbf{0}), \\ \langle \widehat{\mathbf{W}} \widehat{\mathbf{W}}^T, \Delta \Delta^T \rangle &= \text{trace}(\widehat{\mathbf{W}}^T \Delta \Delta^T \widehat{\mathbf{W}}) = 0 \quad (\text{since } \widehat{\mathbf{W}}^T \Delta = \mathbf{0}), \end{aligned}$$

where  $\widehat{\mathbf{W}}^T \mathbf{W} = \mathbf{0}$  since  $\mathbf{U}^T \mathbf{U} - \mathbf{V}^T \mathbf{V} = \mathbf{0}$ ,  $\widehat{\Delta}^T \mathbf{W} = \mathbf{0}$  because  $\widehat{\Delta}^T \mathbf{W} = \alpha \phi_k^T \mathbf{U} - \alpha \psi_k^T \mathbf{V} = \mathbf{0}$  (see (47)), and  $\widehat{\mathbf{W}}^T \Delta = \mathbf{0}$  holds with a similar argument. Plugging these terms into (9) gives

$$\begin{aligned} &[\nabla^2 g(\mathbf{W})](\Delta, \Delta) \\ &= \|\Delta_U \mathbf{V}^T + \mathbf{U} \Delta_V^T\|_F^2 + 2 \langle \mathbf{U} \mathbf{V}^T - \mathbf{X}^*, \Delta_U \Delta_V^T \rangle + \mu \left( \langle \widehat{\mathbf{W}}^T \mathbf{W}, \widehat{\Delta}^T \Delta \rangle + \langle \widehat{\mathbf{W}} \widehat{\Delta}^T, \Delta \mathbf{W}^T \rangle + \langle \widehat{\mathbf{W}} \widehat{\mathbf{W}}^T, \Delta \Delta^T \rangle \right) \\ &= -2\sigma_k. \end{aligned}$$

The proof of the strict saddle property is completed by noting that

$$\|\Delta\|_F^2 = \|\Delta_U\|_F^2 + \|\Delta_V\|_F^2 = \|\phi_k \alpha^T\|_F^2 + \|\psi_k \alpha^T\|_F^2 = 2,$$

which further implies

$$\lambda_{\min}(\nabla^2 g(\mathbf{W})) \leq \frac{[\nabla^2 g(\mathbf{W})](\Delta, \Delta)}{\|\Delta\|_F^2} \leq -\frac{2\sigma_k}{2} = -\|\Lambda - \Sigma\| = -\frac{1}{2} \|\mathbf{W} \mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T}\|,$$

where the first equality holds because

$$\|\Lambda - \Sigma\| = \max_i \sigma_i - \lambda_i = \sigma_k,$$

and the second equality follows since

$$\mathbf{W}\mathbf{W}^T - \mathbf{W}^*\mathbf{W}^{*T} = \frac{1}{2}\mathbf{Q}(\mathbf{\Lambda} - \mathbf{\Sigma})\mathbf{Q}^T, \quad \mathbf{Q} = \begin{bmatrix} \Phi/\sqrt{2} \\ \Psi/\sqrt{2} \end{bmatrix}, \quad \mathbf{Q}^T\mathbf{Q} = \mathbf{I}.$$

We finish the proof of (18) by noting that

$$\sigma_k = \sigma_k(\mathbf{X}^*) \geq \sigma_r(\mathbf{X}^*).$$

Now suppose  $\mathbf{W}^* \in \mathcal{X}$ . Applying (17), which states that the Hessian of  $\rho$  evaluated at any critical point  $\mathbf{W}$  is PSD, we have

$$\begin{aligned} [\nabla^2 g(\mathbf{W}^*)][\Delta, \Delta] &= \left\| \Delta_U \mathbf{V}^{*T} + \mathbf{U}^* \Delta_V^T \right\|_F^2 + 2 \left\langle \mathbf{U}^* \mathbf{V}^{*T} - \mathbf{X}^*, \Delta_U \Delta_V^T \right\rangle + [\nabla^2 \rho(\mathbf{W}^*)][\Delta, \Delta] \\ &\geq \left\| \Delta_U \mathbf{V}^{*T} + \mathbf{U}^* \Delta_V^T \right\|_F^2 + 2 \left\langle \mathbf{U}^* \mathbf{V}^{*T} - \mathbf{X}^*, \Delta_U \Delta_V^T \right\rangle \geq 0 \end{aligned}$$

since  $\mathbf{U}^* \mathbf{V}^{*T} - \mathbf{X}^* = \mathbf{0}$ . We show  $g$  is not strongly convex at  $\mathbf{W}^*$  by arguing that  $\lambda_{\min}(\nabla^2 g(\mathbf{W}^*)) = 0$ . For this purpose, we first recall that  $\mathbf{U}^* = \Phi \mathbf{\Sigma}^{1/2}$ ,  $\mathbf{V}^* = \Psi \mathbf{\Sigma}^{1/2}$ , where we assume  $\mathbf{R} = \mathbf{I}$  without loss of generality. Let  $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_r\}$  be the standard orthobasis for  $\mathbb{R}^r$ , i.e.,  $\mathbf{e}_\ell$  is the  $\ell$ -th column of the  $r \times r$  identity matrix. Construct  $\Delta_{(i,j)} = \begin{bmatrix} \Delta_U^{(i,j)} \\ \Delta_V^{(i,j)} \end{bmatrix}$ , where

$$\Delta_U^{(i,j)} = \mathbf{U}^* \mathbf{e}_j \mathbf{e}_i^T - \mathbf{U}^* \mathbf{e}_i \mathbf{e}_j^T, \quad \Delta_V^{(i,j)} = \mathbf{V}^* \mathbf{e}_j \mathbf{e}_i^T - \mathbf{V}^* \mathbf{e}_i \mathbf{e}_j^T,$$

for any  $1 \leq i < j \leq r$ . That is, the  $\ell$ -th columns of the matrices  $\Delta_U^{(i,j)}$  and  $\Delta_V^{(i,j)}$  are respectively given by

$$\Delta_U^{(i,j)}[:, \ell] = \begin{cases} \sigma_j^{1/2} \phi_j, & \ell = i, \\ -\sigma_i^{1/2} \phi_i, & \ell = j, \\ \mathbf{0}, & \text{otherwise,} \end{cases}, \quad \Delta_V^{(i,j)}[:, \ell] = \begin{cases} \sigma_j^{1/2} \psi_j, & \ell = i, \\ -\sigma_i^{1/2} \psi_i, & \ell = j, \\ \mathbf{0}, & \text{otherwise,} \end{cases}$$

for any  $1 \leq i < j \leq r$ . We then compute the five terms in (9) as follows

$$\begin{aligned} \left\| \Delta_U^{(i,j)} \mathbf{V}^{*T} + \mathbf{U}^* (\Delta_V^{(i,j)})^T \right\|_F^2 &= \left\| \sigma_i^{1/2} \sigma_j^{1/2} (\phi_j \psi_i^T - \phi_i \psi_j^T + \phi_i \psi_j^T - \phi_j \psi_i^T) \right\|_F^2 = 0, \\ \left\langle \mathbf{U}^* \mathbf{V}^{*T} - \mathbf{X}^*, \Delta_U^{(i,j)} (\Delta_V^{(i,j)})^T \right\rangle &= 0 \quad (\text{since } \mathbf{U}^* \mathbf{V}^{*T} - \mathbf{X}^* = \mathbf{0}), \\ \left\langle \widehat{\mathbf{W}}^{*T} \mathbf{W}^*, \widehat{\Delta}_{(i,j)}^T \Delta_{(i,j)} \right\rangle &= 0 \quad (\text{since } \widehat{\mathbf{W}}^{*T} \mathbf{W}^* = \mathbf{0}), \\ \left\langle \widehat{\mathbf{W}}^* \widehat{\Delta}_{(i,j)}^T, \Delta_{(i,j)} \mathbf{W}^{*T} \right\rangle &= \text{trace}(\widehat{\mathbf{W}}^{*T} \Delta_{(i,j)} \mathbf{W}^{*T} \widehat{\Delta}_{(i,j)}) = 0 \quad (\text{since } \widehat{\mathbf{W}}^{*T} \Delta_{(i,j)} = \mathbf{0}), \\ \left\langle \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*T}, \Delta_{(i,j)} \Delta_{(i,j)}^T \right\rangle &= \text{trace}(\widehat{\mathbf{W}}^{*T} \Delta_{(i,j)} \Delta_{(i,j)}^T \widehat{\mathbf{W}}^*) = 0 \quad (\text{since } \widehat{\mathbf{W}}^{*T} \Delta_{(i,j)} = \mathbf{0}), \end{aligned}$$

where  $\widehat{\mathbf{W}}^{*T} \Delta_{(i,j)} = \mathbf{0}$  holds because

$$\widehat{\mathbf{W}}^{*T} \Delta_{(i,j)} = \mathbf{U}^{*T} \mathbf{U}^* (\mathbf{e}_j \mathbf{e}_i^T - \mathbf{e}_i \mathbf{e}_j^T) - \mathbf{V}^{*T} \mathbf{V}^* (\mathbf{e}_j \mathbf{e}_i^T - \mathbf{e}_i \mathbf{e}_j^T)$$

and  $\mathbf{U}^{*T} \mathbf{U}^* = \mathbf{V}^{*T} \mathbf{V}^*$ .

Thus, we obtain the Hessian evaluated at the optimal solution point  $\mathbf{W}^*$  along the direction  $\Delta^{(i,j)}$ :

$$\begin{aligned} &[\nabla^2 g(\mathbf{W}^*)] \left( \Delta^{(i,j)}, \Delta^{(i,j)} \right) \\ &= \left\| \Delta_U^{(i,j)} \mathbf{V}^{*T} + \mathbf{U}^* (\Delta_V^{(i,j)})^T \right\|_F^2 + 2 \left\langle \mathbf{U}^* \mathbf{V}^{*T} - \mathbf{X}^*, \Delta_U^{(i,j)} (\Delta_V^{(i,j)})^T \right\rangle \\ &\quad + \mu \left( \left\langle \widehat{\mathbf{W}}^{*T} \mathbf{W}^*, \widehat{\Delta}_{(i,j)}^T \Delta_{(i,j)} \right\rangle + \left\langle \widehat{\mathbf{W}}^* \widehat{\Delta}_{(i,j)}^T, \Delta_{(i,j)} \mathbf{W}^{*T} \right\rangle + \left\langle \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*T}, \Delta_{(i,j)} \Delta_{(i,j)}^T \right\rangle \right) \\ &= 0 \end{aligned}$$

for all  $1 \leq i < j \leq r$ . This proves that  $g(\mathbf{W})$  is not strongly convex at a global minimum point  $\mathbf{W}^* \in \mathcal{X}$ .

## G Proof of Theorem 2 (strict saddle property of $g(\mathbf{W})$ when over-parameterized)

Let  $\mathbf{X}^* = \Phi \Sigma \Psi^T = \sum_{i=1}^{r'} \sigma_i \phi_i \psi_i^T$  be a reduced SVD of  $\mathbf{X}^*$  with  $r' \leq r$ . Using an approach similar to that in Appendix D for proving Lemma 4, we can show that any  $\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$  is a critical point of  $g(\mathbf{W})$  if and only if  $\mathbf{W} \in \mathcal{C}$  with

$$\mathcal{C} = \left\{ \mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} : \mathbf{U} = \Phi \Lambda^{1/2} \mathbf{R}, \mathbf{V} = \Psi \Lambda^{1/2} \mathbf{R}, \mathbf{R} \mathbf{R}^T = \mathbf{I}_{r'}, \Lambda \text{ is diagonal}, \Lambda \geq \mathbf{0}, (\Sigma - \Lambda) \Sigma = \mathbf{0} \right\}.$$

Recall that

$$\mathcal{X} = \left\{ \mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} : \mathbf{U} = \Phi \Sigma^{1/2} \mathbf{R}, \mathbf{V} = \Psi \Sigma^{1/2} \mathbf{R}, \mathbf{R} \mathbf{R}^T = \mathbf{I}_{r'} \right\}.$$

It is clear that  $\mathcal{X}$  is the set of optimal solutions since for any  $\mathbf{W} \in \mathcal{X}$ ,  $g(\mathbf{W})$  achieves its global minimum, i.e.,  $g(\mathbf{W}) = 0$ .

Using an approach similar to that in Appendix F for proving Theorem 1, we can show that any  $\mathbf{W} \in \mathcal{C} \setminus \mathcal{X}$  is a strict saddle satisfying

$$\lambda_{\min}(\nabla^2 g(\mathbf{W})) \leq -\sigma_{r'}(\mathbf{X}^*).$$

## H Proof of Theorem 3 (strict saddle property of $g(\mathbf{W})$ when under-parameterized)

Let  $\mathbf{X}^* = \Phi \Sigma \Psi^T = \sum_{i=1}^{r'} \sigma_i \phi_i \psi_i^T$  be a reduced SVD of  $\mathbf{X}^*$  with  $r' > r$  and  $\sigma_r(\mathbf{X}^*) > \sigma_{r+1}(\mathbf{X}^*)$ . Using an approach similar to that in Appendix D for proving Lemma 4, we can show that any  $\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$  is a critical point of  $g(\mathbf{W})$  if and only if  $\mathbf{W} \in \mathcal{C}$  with

$$\begin{aligned} \mathcal{C} = \left\{ \mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} : \mathbf{U} = \Phi[:, \Omega] \Lambda^{1/2} \mathbf{R}, \mathbf{V} = \Psi[:, \Omega] \Lambda^{1/2} \mathbf{R}, \right. \\ \left. \Lambda = \Sigma[\Omega, \Omega], \mathbf{R} \mathbf{R}^T = \mathbf{I}_\ell, \Omega \subset \{1, 2, \dots, r'\}, |\Omega| = \ell \leq r \right\}. \end{aligned}$$

Intuitively, a critical point is one such that  $\mathbf{U} \mathbf{V}^T$  is a rank- $\ell$  approximation to  $\mathbf{X}^*$  with  $\ell \leq r$  and  $\mathbf{U}$  and  $\mathbf{V}$  are equal factors of their product  $\mathbf{U} \mathbf{V}^T$ .

It follows from the Eckart-Young-Mirsky theorem [22] that the set of optimal solutions is given by

$$\mathcal{X} = \left\{ \mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} : \mathbf{U} = \Phi[:, 1:r] \Lambda^{1/2} \mathbf{R}, \mathbf{V} = \Psi[:, 1:r] \Lambda^{1/2} \mathbf{R}, \Lambda = \Sigma[1:r, 1:r], \mathbf{R} \in \mathcal{O}_r \right\}.$$

Now we characterize any  $\mathbf{W} \in \mathcal{C} \setminus \mathcal{X}$  by letting  $\mathbf{W} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$ , where

$$\begin{aligned} \mathbf{U} &= \Phi[:, \Omega] \Lambda^{1/2} \mathbf{R}, \mathbf{V} = \Psi[:, \Omega] \Lambda^{1/2} \mathbf{R}, \\ \Lambda &= \Sigma[\Omega, \Omega], \mathbf{R} \in \mathbb{R}^{\ell \times r}, \mathbf{R} \mathbf{R}^T = \mathbf{I}_\ell, \Omega \subset \{1, 2, \dots, r'\}, |\Omega| = \ell \leq r, \Omega \neq \{1, 2, \dots, r\}. \end{aligned}$$

Let  $\alpha \in \mathbb{R}^r$  be the eigenvector associated with the smallest eigenvalue of  $\mathbf{U}^T \mathbf{U}$  (or  $\mathbf{V}^T \mathbf{V}$ ). By the typical structures in  $\mathbf{U}$  and  $\mathbf{V}$  (see the above equation), we have

$$\|\mathbf{V} \alpha\|_F^2 = \|\mathbf{U} \alpha\|_F^2 = \sigma_r^2(\mathbf{U}) = \begin{cases} \sigma_j(\mathbf{X}^*), & |\Omega| = r \text{ and } j = \max \Omega \\ 0, & |\Omega| < r, \end{cases} \quad (49)$$

where  $j > r$  because  $\Omega \neq \{1, 2, \dots, r\}$ . Note that there always exists an index

$$i \in \{1, 2, \dots, r\}, i \neq \Omega$$

since  $\Omega \neq \{1, 2, \dots, r\}$  and  $|\Omega| \leq r$ . We construct  $\Delta$  by setting

$$\Delta_U = \phi_i \alpha^T, \quad \Delta_V = \psi_i \alpha^T.$$

Since  $i \notin \Omega$ , we have

$$\begin{aligned} U^T \Delta_U &= U^T \phi_i \alpha^T = \mathbf{0}, \\ V^T \Delta_V &= V^T \psi_i \alpha^T = \mathbf{0}. \end{aligned} \tag{50}$$

We compute the five terms in (9) as follows

$$\begin{aligned} \|\Delta_U V^T + U \Delta_V^T\|_F^2 &= \|\Delta_U V^T\|_F^2 + \|U \Delta_V^T\|_F^2 + 2 \text{trace}(U^T \Delta_U V^T \Delta_V) = 2\sigma_r^2(U), \\ \langle UV^T - X^*, \Delta_U \Delta_V^T \rangle &= \langle UV^T - X^*, \phi_i \psi_i^T \rangle = -\langle X^*, \phi_i \psi_i^T \rangle = -\sigma_i(X^*), \\ \langle \widehat{W}^T W, \widehat{\Delta}^T \Delta \rangle &= 0 \quad (\text{since } \widehat{W}^T W = \mathbf{0}), \\ \langle \widehat{W} \widehat{\Delta}^T, \Delta W^T \rangle &= \text{trace}(\widehat{W}^T \Delta W^T \widehat{\Delta}) = 0 \quad (\text{since } \widehat{W}^T \Delta = \mathbf{0}), \\ \langle \widehat{W} \widehat{W}^T, \Delta \Delta^T \rangle &= \text{trace}(\widehat{W}^T \Delta \Delta^T \widehat{W}) = 0 \quad (\text{since } \widehat{W}^T \Delta = \mathbf{0}), \end{aligned}$$

where the last equality in the first line holds because  $U^T \Delta_U = \mathbf{0}$  (see (50)) and  $\|\Delta_U V^T\|_F^2 = \|U \Delta_V^T\|_F^2 = \sigma_r^2(U)$  (see (49)),  $\widehat{W}^T W = \mathbf{0}$  in the third line holds since  $U^T U - V^T V = \mathbf{0}$ , and  $\widehat{W}^T \Delta = \mathbf{0}$  in the fourth and last lines holds because

$$\widehat{W}^T \Delta = U^T \Delta_U - V^T \Delta_V = \mathbf{0}.$$

Now plugging these terms into (9) yields

$$\begin{aligned} &[\nabla^2 g(W)](\Delta, \Delta) \\ &= \|\Delta_U V^T + U \Delta_V^T\|_F^2 + 2 \langle UV^T - X^*, \Delta_U \Delta_V^T \rangle + \mu (\langle \widehat{W}^T W, \widehat{\Delta}^T \Delta \rangle + \langle \widehat{W} \widehat{\Delta}^T, \Delta W^T \rangle + \langle \widehat{W} \widehat{W}^T, \Delta \Delta^T \rangle) \\ &= -2(\sigma_i(X^*) - \sigma_r^2(U)). \end{aligned}$$

The proof of the strict saddle property is completed by noting that

$$\|\Delta\|_F^2 = \|\Delta_U\|_F^2 + \|\Delta_V\|_F^2 = 2,$$

which further implies

$$\lambda_{\min}(\nabla^2 g(W)) \leq -2 \frac{\sigma_i(X^*) - \sigma_r^2(U)}{\|\Delta\|_F^2} = -(\sigma_i(X^*) - \sigma_r^2(U)) \leq -(\sigma_r(X^*) - \sigma_{r+1}(X^*)),$$

where the last inequality holds because of (49) and because  $i \leq r$ .

## I Proof of Theorem 4 (robust strict saddle for $g(W)$ )

We first establish the following useful results.

*Lemma 7.* For any two PSD matrices  $A, B \in \mathbb{R}^{n \times n}$ , we have

$$\sigma_n(A) \text{trace}(B) \leq \text{trace}(AB) \leq \|A\| \text{trace}(B).$$

*Proof of Lemma 7.* Let  $A = \Phi_1 \Lambda_1 \Phi_1^T$  and  $B = \Phi_2 \Lambda_2 \Phi_2^T$  be the eigendecompositions of  $A$  and  $B$ , respectively. Here  $\Lambda_1$  ( $\Lambda_2$ ) is a diagonal matrix with the eigenvalues of  $A$  ( $B$ ) along its diagonal. We first rewrite  $\text{trace}(AB)$  as

$$\text{trace}(AB) = \text{trace}(\Lambda_1 \Phi_1^T \Phi_2 \Lambda_2 \Phi_2^T \Phi_1).$$

Noting that  $\Lambda_1$  is a diagonal matrix, we have

$$\text{trace}(\Lambda_1 \Phi_1^T \Phi_2 \Lambda_2 \Phi_2^T \Phi_1) \geq \min_i \Lambda_1[i, i] \cdot \text{trace}(\Phi_1^T \Phi_2 \Lambda_2 \Phi_2^T \Phi_1) = \sigma_n(A) \text{trace}(B).$$

The other direction follows similarly. □

Corollary 1. For any two matrices  $\mathbf{A} \in \mathbb{R}^{r \times r}$  and  $\mathbf{B} \in \mathbb{R}^{n \times r}$ , we have

$$\sigma_r(\mathbf{A})\|\mathbf{B}\|_F \leq \|\mathbf{AB}\|_F \leq \|\mathbf{A}\|\|\mathbf{B}\|_F.$$

We provide one more result before proceeding to prove the main theorem.

Lemma 8. Suppose  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times r}$  such that  $\mathbf{A}^T \mathbf{B} = \mathbf{B}^T \mathbf{A} \succeq \mathbf{0}$  is PSD. If  $\|\mathbf{A} - \mathbf{B}\| \leq \frac{\sqrt{2}}{2}\sigma_r(\mathbf{B})$ , we have

$$\underbrace{\left\langle \left( \mathbf{AA}^T - \mathbf{BB}^T \right) \mathbf{A}, \mathbf{A} - \mathbf{B} \right\rangle}_{(\aleph_1)} \geq \frac{1}{16} \underbrace{\text{trace} \left( (\mathbf{A} - \mathbf{B})^T (\mathbf{A} - \mathbf{B}) \mathbf{B}^T \mathbf{B} \right)}_{(\aleph_2)} + \frac{1}{16} \underbrace{\left\| \mathbf{AA}^T - \mathbf{BB}^T \right\|_F^2}_{(\aleph_3)}. \quad (51)$$

*Proof.* Denote  $\mathbf{E} = \mathbf{A} - \mathbf{B}$ . We first rewrite the terms  $(\aleph_1)$ ,  $(\aleph_2)$  and  $(\aleph_3)$  as follows

$$\begin{aligned} (\aleph_1) &= \text{trace} \left( \left( \mathbf{E}^T \mathbf{E} \right)^2 + 3\mathbf{E}^T \mathbf{E} \mathbf{E}^T \mathbf{B} + \left( \mathbf{E}^T \mathbf{B} \right)^2 + \mathbf{E}^T \mathbf{E} \mathbf{B}^T \mathbf{B} \right), \\ (\aleph_2) &= \text{trace} \left( \mathbf{E}^T \mathbf{E} \mathbf{B}^T \mathbf{B} \right), \\ (\aleph_3) &= \text{trace} \left( \left( \mathbf{E}^T \mathbf{E} \right)^2 + 4\mathbf{E}^T \mathbf{E} \mathbf{E}^T \mathbf{B} + 2 \left( \mathbf{E}^T \mathbf{B} \right)^2 + 2\mathbf{E}^T \mathbf{E} \mathbf{B}^T \mathbf{B} \right), \end{aligned}$$

where  $\mathbf{E}^T \mathbf{B} = \mathbf{A}^T \mathbf{B} - \mathbf{B}^T \mathbf{B} = \mathbf{B}^T \mathbf{E}$ . Now we have

$$\begin{aligned} (\aleph_1) - \frac{1}{16}(\aleph_2) - \frac{1}{16}(\aleph_3) &= \text{trace} \left( \frac{15}{16} \left( \mathbf{E}^T \mathbf{E} \right)^2 + \frac{11}{4} \mathbf{E}^T \mathbf{E} \mathbf{E}^T \mathbf{B} + \frac{7}{8} \left( \mathbf{E}^T \mathbf{B} \right)^2 + \frac{13}{16} \mathbf{E}^T \mathbf{E} \mathbf{B}^T \mathbf{B} \right) \\ &= \left\| \sqrt{\frac{121}{56}} \mathbf{E}^T \mathbf{E} + \sqrt{\frac{7}{8}} \mathbf{E}^T \mathbf{B} \right\|_F^2 + \text{trace} \left( \frac{13}{16} \mathbf{E}^T \mathbf{E} \mathbf{B}^T \mathbf{B} - \frac{137}{112} \mathbf{E}^T \mathbf{E} \mathbf{E}^T \mathbf{E} \right) \\ &\geq \text{trace} \left( \frac{13}{16} \mathbf{E}^T \mathbf{E} \sigma_r^2(\mathbf{B}) - \frac{137}{112} \mathbf{E}^T \mathbf{E} \|\mathbf{E}\|^2 \right) \\ &\geq \text{trace} \left( \left( \frac{13}{16} - \frac{137}{112} \frac{1}{2} \right) \sigma_r^2(\mathbf{B}) \mathbf{E}^T \mathbf{E} \right) \\ &\geq 0, \end{aligned}$$

where the third line follows from Lemma 7 and the fourth line holds because by assumption  $\|\mathbf{E}\| \leq \frac{\sqrt{2}}{2}\sigma_r(\mathbf{B})$ .  $\square$

Now we turn to prove the main results. Recall that  $\mu = \frac{1}{2}$  throughout the proof.

## I.1 Regularity condition for the region $\mathcal{R}_1$

It follows from Lemma 6 that  $\mathbf{W}^T \mathbf{W}^* \mathbf{R} = \mathbf{R}^T \mathbf{W}^{*T} \mathbf{W}$  is PSD, where  $\mathbf{R} = \arg \min_{\mathbf{R}' \in \mathcal{O}_r} \|\mathbf{W} - \mathbf{W}^* \mathbf{R}'\|_F^2$ . We first perform the change of variable  $\mathbf{W}^* \mathbf{R} \rightarrow \mathbf{W}^*$  to avoid  $\mathbf{R}$  in the following equations. With this change of variable we have instead  $\mathbf{W}^T \mathbf{W}^* = \mathbf{W}^{*T} \mathbf{W}$  is PSD. We now rewrite the gradient  $\nabla g(\mathbf{W})$  as follows:

$$\begin{aligned} \nabla g(\mathbf{W}) &= \begin{bmatrix} \mathbf{0} & \mathbf{U} \mathbf{V}^T - \mathbf{U}^* \mathbf{V}^{*T} \\ \mathbf{V} \mathbf{U}^T - \mathbf{V}^* \mathbf{U}^{*T} & \mathbf{0} \end{bmatrix} \mathbf{W} + \mu \widehat{\mathbf{W}} (\widehat{\mathbf{W}}^T \mathbf{W}) \\ &= \frac{1}{2} \left( \mathbf{W} \mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T} \right) \mathbf{W} + \frac{1}{2} \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*T} \mathbf{W} + \left( \mu - \frac{1}{2} \right) \widehat{\mathbf{W}} \widehat{\mathbf{W}}^T \mathbf{W} \\ &= \frac{1}{2} \left( \mathbf{W} \mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T} \right) \mathbf{W} + \frac{1}{2} \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*T} \mathbf{W}. \end{aligned} \quad (52)$$

Plugging this into the left hand side of (21) gives

$$\begin{aligned} \langle \nabla g(\mathbf{W}), \mathbf{W} - \mathbf{W}^* \rangle &= \frac{1}{2} \left\langle \left( \mathbf{W} \mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T} \right) \mathbf{W}, \mathbf{W} - \mathbf{W}^* \right\rangle + \frac{1}{2} \left\langle \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*T} \mathbf{W}, \mathbf{W} - \mathbf{W}^* \right\rangle \\ &= \frac{1}{2} \left\langle \left( \mathbf{W} \mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T} \right) \mathbf{W}, \mathbf{W} - \mathbf{W}^* \right\rangle + \frac{1}{2} \left\langle \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*T}, \mathbf{W} \mathbf{W}^T \right\rangle \end{aligned} \quad (53)$$



where the last line follows from the fact that  $\mathbf{W}^{*\text{T}}\widehat{\mathbf{W}}^* = \mathbf{0}$ . We first show the first term in the right hand side of the above equation is sufficiently large

$$\begin{aligned}
& \left\langle (\mathbf{W}\mathbf{W}^{\text{T}} - \mathbf{W}^*\mathbf{W}^{*\text{T}}) \mathbf{W}, \mathbf{W} - \mathbf{W}^* \right\rangle \\
& \geq \frac{1}{16} \text{trace} \left( (\mathbf{W} - \mathbf{W}^*)^{\text{T}} (\mathbf{W} - \mathbf{W}^*) \mathbf{W}^{*\text{T}} \mathbf{W}^* \right) + \frac{1}{16} \left\| \mathbf{W}\mathbf{W}^{\text{T}} - \mathbf{W}^*\mathbf{W}^{*\text{T}} \right\|_F^2 \\
& \geq \frac{1}{16} \sigma_r(\mathbf{W}^{*\text{T}} \mathbf{W}^*) \left\| \mathbf{W} - \mathbf{W}^* \right\|_F^2 + \frac{1}{16} \left\| \mathbf{W}\mathbf{W}^{\text{T}} - \mathbf{W}^*\mathbf{W}^{*\text{T}} \right\|_F^2 \\
& = \frac{1}{8} \sigma_r(\mathbf{X}^*) \left\| \mathbf{W} - \mathbf{W}^* \right\|_F^2 + \frac{1}{16} \left\| \mathbf{W}\mathbf{W}^{\text{T}} - \mathbf{W}^*\mathbf{W}^{*\text{T}} \right\|_F^2,
\end{aligned} \tag{54}$$

where the first inequality follows from Lemma 8 since  $\mathbf{W}^{\text{T}} \mathbf{W}^* = \mathbf{W}^{*\text{T}} \mathbf{W}$  is PSD and  $\left\| \mathbf{W} - \mathbf{W}^* \right\| \leq \sigma_r^{1/2}(\mathbf{X}^*) = \frac{\sqrt{2}}{2} \sigma_r(\mathbf{W}^*)$ , the second inequality follows from Lemma 7, and the last line holds because  $\sigma_r(\widehat{\mathbf{W}}^{*\text{T}} \widehat{\mathbf{W}}^*) = \sigma_r(\widehat{\mathbf{U}}^{*\text{T}} \widehat{\mathbf{U}}^* + \widehat{\mathbf{V}}^{*\text{T}} \widehat{\mathbf{V}}^*) = 2\sigma_r(\mathbf{\Sigma}) = 2\sigma_r(\mathbf{X}^*)$ . We then show the second term in the right hand side of (53) is lower bounded by

$$\begin{aligned}
\left\langle \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*\text{T}}, \mathbf{W}\mathbf{W}^{\text{T}} \right\rangle &= \frac{1}{2 \left\| \mathbf{X}^* \right\|} \left\| \widehat{\mathbf{W}}^{*\text{T}} \widehat{\mathbf{W}}^* \right\| \text{trace} \left( \widehat{\mathbf{W}}^{*\text{T}} \mathbf{W}\mathbf{W}^{\text{T}} \widehat{\mathbf{W}}^* \right) \\
&\geq \frac{1}{2 \left\| \mathbf{X}^* \right\|} \text{trace} \left( \widehat{\mathbf{W}}^{*\text{T}} \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*\text{T}} \mathbf{W}\mathbf{W}^{\text{T}} \widehat{\mathbf{W}}^* \right) \\
&= \frac{1}{2 \left\| \mathbf{X}^* \right\|} \left\| \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*\text{T}} \mathbf{W} \right\|_F^2
\end{aligned} \tag{55}$$

where the first line holds because  $\left\| \widehat{\mathbf{W}}^{*\text{T}} \widehat{\mathbf{W}}^* \right\| = \left\| \widehat{\mathbf{U}}^{*\text{T}} \widehat{\mathbf{U}}^* + \widehat{\mathbf{V}}^{*\text{T}} \widehat{\mathbf{V}}^* \right\| = 2 \left\| \mathbf{\Sigma} \right\| = 2 \left\| \mathbf{X}^* \right\|$ , and the inequality follows from Lemma 7.

On the other hand, we attempt to control the gradient of  $g(\mathbf{W})$ . To that end, it follows from (52) that

$$\begin{aligned}
\left\| \nabla g(\mathbf{W}) \right\|_F^2 &= \frac{1}{4} \left\| (\mathbf{W}\mathbf{W}^{\text{T}} - \mathbf{W}^*\mathbf{W}^{*\text{T}}) \mathbf{W} + \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*\text{T}} \mathbf{W} \right\|_F^2 \\
&\leq \frac{1}{4} \left( \frac{48}{47} \left\| (\mathbf{W}\mathbf{W}^{\text{T}} - \mathbf{W}^*\mathbf{W}^{*\text{T}}) \mathbf{W} \right\|_F^2 + 48 \left\| \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*\text{T}} \mathbf{W} \right\|_F^2 \right) \\
&\leq \frac{1}{4} \left( \frac{48}{47} \left\| \mathbf{W} \right\|^2 \left\| \mathbf{W}\mathbf{W}^{\text{T}} - \mathbf{W}^*\mathbf{W}^{*\text{T}} \right\|_F^2 + 48 \left\| \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*\text{T}} \mathbf{W} \right\|_F^2 \right),
\end{aligned} \tag{56}$$

where the first inequality holds since  $(a+b)^2 \leq \frac{1+\epsilon}{\epsilon} a^2 + (1+\epsilon) b^2$  for any  $\epsilon > 0$ .

Combining (53)-(56), we can conclude the proof of (21) as long as we can show the following inequality:

$$\frac{1}{8} \left\| \mathbf{W}\mathbf{W}^{\text{T}} - \mathbf{W}^*\mathbf{W}^{*\text{T}} \right\|_F^2 \geq \frac{1}{47} \frac{\left\| \mathbf{W} \right\|^2}{\left\| \mathbf{X}^* \right\|^2} \left\| \mathbf{W}\mathbf{W}^{\text{T}} - \mathbf{W}^*\mathbf{W}^{*\text{T}} \right\|_F^2.$$

To that end, we upper bound  $\left\| \mathbf{W} \right\|$  as follows:

$$\begin{aligned}
\left\| \mathbf{W} \right\| &\leq \left\| \mathbf{W}^* \right\| + \left\| \mathbf{W} - \mathbf{W}^* \right\| \\
&\leq \sqrt{2} \sigma_1(\mathbf{X}^*) + \left\| \mathbf{W} - \mathbf{W}^* \right\|_F \\
&\leq (\sqrt{2} + 1) \sigma_1(\mathbf{X}^*)
\end{aligned}$$

since  $\left\| \mathbf{W}^* \right\| = \sqrt{2} \sigma_1(\mathbf{X}^*)$  and  $\text{dist}(\mathbf{W}, \mathbf{W}^*) \leq \sigma_r(\mathbf{X}^*)$ . This completes the proof of (21).

## I.2 Negative curvature for the region $\mathcal{R}_2$

To show (22), we utilize a strategy similar to that used in Appendix F for proving the strict saddle property of  $g(\mathbf{W})$  by constructing a direction  $\mathbf{\Delta}$  such that the Hessian evaluated at  $\mathbf{W}$  along this direction is negative. For this purpose, denote

$$\mathbf{Q} = \begin{bmatrix} \mathbf{\Phi}/\sqrt{2} \\ \mathbf{\Psi}/\sqrt{2} \end{bmatrix}, \tag{57}$$

where we recall that  $\Phi$  and  $\Psi$  consist of the left and right singular vectors of  $\mathbf{X}^*$ , respectively. The optimal solution  $\mathbf{W}^*$  has a compact SVD  $\mathbf{W}^* = \mathbf{Q}(\sqrt{2\mathbf{\Sigma}^{1/2}})\mathbf{R}$ . For notational convenience, we denote  $\bar{\mathbf{\Sigma}} = 2\mathbf{\Sigma}$ , where  $\bar{\mathbf{\Sigma}}$  is a diagonal matrix whose diagonal entries in the upper left corner are  $\bar{\sigma}_1, \dots, \bar{\sigma}_r$ .

For any  $\mathbf{W}$ , we can always divide it into two parts, the projections onto the column spaces of  $\mathbf{Q}$  and its orthogonal complement, respectively. Equivalently, we can write

$$\mathbf{W} = \mathbf{Q}\bar{\mathbf{\Lambda}}^{1/2}\mathbf{R} + \mathbf{E}, \quad (58)$$

where  $\mathbf{Q}\bar{\mathbf{\Lambda}}^{1/2}\mathbf{R}$  is a compact SVD form representing the projection of  $\mathbf{W}$  onto the column space of  $\mathbf{Q}$ , and  $\mathbf{E}^T\mathbf{Q} = \mathbf{0}$  (i.e.,  $\mathbf{E}$  is orthogonal to  $\mathbf{Q}$ ). Here  $\mathbf{R} \in \mathcal{O}_r$  and  $\bar{\mathbf{\Lambda}}$  is a diagonal matrix whose diagonal entries in the upper left corner are  $\bar{\lambda}_1, \dots, \bar{\lambda}_r$ , but the diagonal entries are not necessarily placed either in decreasing or increasing order. In order to characterize the neighborhood near all strict saddles  $\mathcal{C} \setminus \mathcal{X}$ , we consider  $\mathbf{W}$  such that  $\sigma_r(\mathbf{W}) \leq \sqrt{\frac{3}{8}}\sigma_r^{1/2}(\mathbf{X}^*)$ . Let  $k := \arg \min_i \bar{\lambda}_i$  denote the location of the smallest diagonal entry in  $\bar{\mathbf{\Lambda}}$ . It is clear that

$$\bar{\lambda}_k \leq \sigma_r^2(\mathbf{W}) \leq \frac{3}{8}\sigma_r(\mathbf{X}^*). \quad (59)$$

Let  $\alpha \in \mathbb{R}^r$  be the eigenvector associated with the smallest eigenvalue of  $\mathbf{W}^T\mathbf{W}$ .

Recall that  $\mu = \frac{1}{2}$ . We show that the function  $g(\mathbf{W})$  at  $\mathbf{W}$  has directional negative curvature along the direction

$$\Delta = \mathbf{q}_k \alpha^T. \quad (60)$$

We repeat the Hessian evaluated at  $\mathbf{W}$  for  $\Delta$  as follows

$$\begin{aligned} [\nabla^2 g(\mathbf{W})](\Delta, \Delta) &= \underbrace{\left\| \Delta_U \mathbf{V}^T + \mathbf{U} \Delta_V^T \right\|_F^2}_{\Pi_1} + 2 \underbrace{\left\langle \mathbf{U} \mathbf{V}^T - \mathbf{X}^*, \Delta_U \Delta_V^T \right\rangle}_{\Pi_2} \\ &\quad + \frac{1}{2} \left( \underbrace{\left\langle \widehat{\Delta} \widehat{\mathbf{W}}^T, \Delta \mathbf{W}^T \right\rangle}_{\Pi_3} + \underbrace{\left\langle \widehat{\mathbf{W}} \widehat{\Delta}^T, \Delta \mathbf{W}^T \right\rangle}_{\Pi_4} + \underbrace{\left\langle \widehat{\mathbf{W}} \widehat{\mathbf{W}}^T, \Delta \Delta^T \right\rangle}_{\Pi_5} \right). \end{aligned}$$

The remaining part is to bound the five terms.

**Bounding terms  $\Pi_1$ ,  $\Pi_3$  and  $\Pi_4$ :** We first rewrite these three terms:

$$\begin{aligned} \Pi_1 &= \|\Delta_U \mathbf{V}^T\|_F^2 + \|\mathbf{U} \Delta_V^T\|_F^2 + 2 \left\langle \mathbf{U} \Delta_V^T, \Delta_U \mathbf{V}^T \right\rangle, \\ \Pi_3 &= \left\langle \widehat{\Delta} \widehat{\mathbf{W}}^T, \Delta \mathbf{W}^T \right\rangle = \|\Delta_U \mathbf{U}^T\|_F^2 + \|\Delta_V \mathbf{V}^T\|_F^2 - \|\Delta_U \mathbf{V}^T\|_F^2 - \|\Delta_V \mathbf{U}^T\|_F^2, \\ \Pi_4 &= \left\langle \mathbf{U} \Delta_U^T, \Delta_U \mathbf{U}^T \right\rangle + \left\langle \mathbf{V} \Delta_V^T, \Delta_V \mathbf{V}^T \right\rangle - 2 \left\langle \mathbf{U} \Delta_V^T, \Delta_U \mathbf{V}^T \right\rangle \\ &\leq \|\Delta_U \mathbf{U}^T\|_F^2 + \|\Delta_V \mathbf{V}^T\|_F^2 - 2 \left\langle \mathbf{U} \Delta_V^T, \Delta_U \mathbf{V}^T \right\rangle, \end{aligned}$$

which implies

$$\begin{aligned} \Pi_1 + \frac{1}{2}\Pi_3 + \frac{1}{2}\Pi_4 &\leq \|\Delta_U \mathbf{V}^T\|_F^2 + \|\mathbf{U} \Delta_V^T\|_F^2 + \|\Delta_U \mathbf{U}^T\|_F^2 + \|\Delta_V \mathbf{V}^T\|_F^2 \\ &\quad - \frac{1}{2}\|\Delta_U \mathbf{V}^T\|_F^2 - \frac{1}{2}\|\Delta_V \mathbf{U}^T\|_F^2 + \left\langle \mathbf{U} \Delta_V^T, \Delta_U \mathbf{V}^T \right\rangle \\ &= \|\mathbf{W} \Delta^T\|_F^2 - \frac{1}{2} \left\| \Delta_U \mathbf{V}^T - \mathbf{U} \Delta_V^T \right\|_F^2 \\ &\leq \|\mathbf{W} \Delta^T\|_F^2. \end{aligned} \quad (61)$$

Noting that  $\Delta^T \Delta = \alpha \mathbf{q}_k^T \mathbf{q}_k \alpha^T = \alpha \alpha^T$ , we now compute  $\|\mathbf{W} \Delta^T\|_F^2$  as

$$\|\mathbf{W} \Delta^T\|_F^2 = \text{trace}(\mathbf{W}^T \mathbf{W} \Delta^T \Delta) = \text{trace}(\mathbf{W}^T \mathbf{W} \alpha \alpha^T) = \sigma_r^2(\mathbf{W}).$$

Plugging this into (61) gives

$$\Pi_1 + \frac{1}{2}\Pi_3 + \frac{1}{2}\Pi_4 \leq \sigma_r^2(\mathbf{W}). \quad (62)$$

**Bounding terms  $\Pi_2$  and  $\Pi_5$ :** To obtain an upper bound for the term  $\Pi_2$ , we first rewrite it as follows

$$\begin{aligned} \Pi_2 &= \left\langle \mathbf{U}\mathbf{V}^T - \mathbf{X}^*, \Delta_U \Delta_V^T \right\rangle = \frac{1}{2} \left\langle \begin{bmatrix} \mathbf{0} & \mathbf{U}\mathbf{V}^T - \mathbf{U}^* \mathbf{V}^{*T} \\ \mathbf{V}\mathbf{U}^T - \mathbf{V}^* \mathbf{U}^{*T} & \mathbf{0} \end{bmatrix}, \Delta \Delta^T \right\rangle \\ &= \frac{1}{4} \left\langle \mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T}, \Delta \Delta^T \right\rangle - \frac{1}{4} \left\langle \widehat{\mathbf{W}} \widehat{\mathbf{W}}^T \Delta \Delta^T \right\rangle + \frac{1}{4} \left\langle \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*T}, \Delta \Delta^T \right\rangle. \end{aligned}$$

We then have

$$2\Pi_2 + \frac{1}{2}\Pi_5 = \frac{1}{2} \left\langle \mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T}, \Delta \Delta^T \right\rangle + \frac{1}{2} \left\langle \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*T}, \Delta \Delta^T \right\rangle. \quad (63)$$

To bound these two terms in the above equation, we note that

$$\Delta \Delta^T = \sum_{i=1}^r \alpha_i^2 \mathbf{q}_i \mathbf{q}_i^T = \mathbf{q}_k \mathbf{q}_k^T = \frac{1}{2} \begin{bmatrix} \phi_k \phi_k^T & \phi_k \psi_k^T \\ \psi_k \phi_k^T & \psi_k \psi_k^T \end{bmatrix}.$$

Then we have

$$\left\langle \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*T}, \Delta \Delta^T \right\rangle = \frac{1}{2} \left\langle \begin{bmatrix} \Phi \Sigma \Phi^T & -\Phi \Sigma \Psi^T \\ -\Psi \Sigma \Phi^T & \Psi \Sigma \Psi^T \end{bmatrix}, \begin{bmatrix} \phi_k \phi_k^T & \phi_k \psi_k^T \\ \psi_k \phi_k^T & \psi_k \psi_k^T \end{bmatrix} \right\rangle = 0,$$

and

$$\begin{aligned} \left\langle \mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T}, \Delta \Delta^T \right\rangle &= \left\langle \mathbf{Q} \bar{\Lambda} \mathbf{Q}^T - 2\mathbf{Q} \bar{\Lambda}^{1/2} \mathbf{R} \mathbf{E}^T + \mathbf{E} \mathbf{E}^T - \mathbf{Q} \bar{\Sigma} \mathbf{Q}^T, \mathbf{q}_k \mathbf{q}_k^T \right\rangle \\ &= \bar{\lambda}_k - \bar{\sigma}_k \end{aligned}$$

where the last utilizes the fact that  $\mathbf{E}^T \mathbf{q}_k = \mathbf{0}$  since  $\mathbf{E}$  is orthogonal to  $\mathbf{Q}$ .

Plugging these into (63) gives

$$2\Pi_2 + \frac{1}{2}\Pi_5 = \frac{1}{2}(\bar{\lambda}_k - \bar{\sigma}_k). \quad (64)$$

**Merging together:** Putting (62) and (64) together yields

$$\begin{aligned} [\nabla^2 g(\mathbf{W})](\Delta, \Delta) &= \Pi_1 + \frac{1}{2}\Pi_3 + \frac{1}{2}\Pi_4 + 2\Pi_2 + \frac{1}{2}\Pi_5 \\ &\leq \sigma_r^2(\mathbf{W}) + \frac{1}{2}(\bar{\lambda}_k - \bar{\sigma}_k) \\ &\leq \frac{1}{2}\sigma_r(\mathbf{X}^*) + \frac{1}{2}\left(\frac{1}{2}\sigma_r(\mathbf{X}^*) - 2\sigma_r(\mathbf{X}^*)\right) \\ &\leq -\frac{1}{4}\sigma_r(\mathbf{X}^*), \end{aligned}$$

where the third line follows because by assumption  $\sigma_r(\mathbf{W}) \leq \sqrt{\frac{1}{2}\sigma_r^{1/2}(\mathbf{X}^*)}$ , by construction  $\bar{\lambda}_k \leq \frac{1}{2}\sigma_r(\mathbf{X}^*)$  (see (59)), and  $\bar{\sigma}_k \geq \bar{\sigma}_r = 2\sigma_r(\mathbf{X}^*)$ . This completes the proof of (22).

### I.3 Large gradient for the region $\mathcal{R}'_3 \cup \mathcal{R}''_3 \cup \mathcal{R}'''_3$ :

In order to show that  $g(\mathbf{W})$  has a large gradient in the three regions  $\mathcal{R}'_3 \cup \mathcal{R}''_3 \cup \mathcal{R}'''_3$ , we first provide a lower bound for the gradient. By (52), we have

$$\begin{aligned}
\|\nabla g(\mathbf{W})\|_F^2 &= \frac{1}{4} \left\| \left( \mathbf{W}\mathbf{W}^T - \mathbf{W}^*\mathbf{W}^{*\top} \right) \mathbf{W} + \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*\top} \mathbf{W} \right\|_F^2 \\
&= \frac{1}{4} \left( \left\| \left( \mathbf{W}\mathbf{W}^T - \mathbf{W}^*\mathbf{W}^{*\top} \right) \mathbf{W} \right\|_F^2 + \left\| \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*\top} \mathbf{W} \right\|_F^2 \right) \\
&\quad + \frac{1}{2} \left\langle \left( \mathbf{W}\mathbf{W}^T - \mathbf{W}^*\mathbf{W}^{*\top} \right) \mathbf{W}, \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*\top} \mathbf{W} \right\rangle \\
&= \frac{1}{4} \left( \left\| \left( \mathbf{W}\mathbf{W}^T - \mathbf{W}^*\mathbf{W}^{*\top} \right) \mathbf{W} \right\|_F^2 + \left\| \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*\top} \mathbf{W} \right\|_F^2 \right) \\
&\quad + \frac{1}{2} \left\langle \mathbf{W}\mathbf{W}^T \mathbf{W}\mathbf{W}^T, \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*\top} \right\rangle \\
&\geq \frac{1}{4} \left\| \left( \mathbf{W}\mathbf{W}^T - \mathbf{W}^*\mathbf{W}^{*\top} \right) \mathbf{W} \right\|_F^2,
\end{aligned} \tag{65}$$

where the third equality follows because  $\mathbf{W}^{*\top} \widehat{\mathbf{W}}^* = \mathbf{U}^{*\top} \mathbf{U}^* - \mathbf{V}^{*\top} \mathbf{V}^* = \mathbf{0}$  and the last line utilizes the fact that the inner product between two PSD matrices is nonnegative.

#### I.3.1 Large gradient for the region $\mathcal{R}'_3$ :

To show  $\|\nabla g(\mathbf{W})\|_F^2$  is large for any  $\mathbf{W} \in \mathcal{R}'_3$ , again, for any  $\mathbf{W} \in \mathbb{R}^{(n+m) \times r}$ , we utilize (58) to write  $\mathbf{W} = \mathbf{Q} \bar{\mathbf{\Lambda}}^{1/2} \mathbf{R} + \mathbf{E}$ , where  $\mathbf{Q}$  is defined in (57),  $\mathbf{Q} \bar{\mathbf{\Lambda}}^{1/2} \mathbf{R}$  is a compact SVD form representing the projection of  $\mathbf{W}$  onto the column space of  $\mathbf{Q}$ , and  $\mathbf{E}^\top \mathbf{Q} = \mathbf{0}$  (i.e.,  $\mathbf{E}$  is orthogonal to  $\mathbf{Q}$ ). Plugging this form of  $\mathbf{W}$  into the last term of (65) gives

$$\begin{aligned}
&\left\| \left( \mathbf{W}\mathbf{W}^T - \mathbf{W}^*\mathbf{W}^{*\top} \right) \mathbf{W} \right\|_F^2 \\
&= \left\| \mathbf{Q} \bar{\mathbf{\Lambda}}^{1/2} (\bar{\mathbf{\Lambda}} - \bar{\mathbf{\Sigma}}) \mathbf{R} + \mathbf{Q} \bar{\mathbf{\Lambda}}^{1/2} \mathbf{R} \mathbf{E} \mathbf{E}^\top + \mathbf{E} \mathbf{R}^\top \bar{\mathbf{\Lambda}} \mathbf{R} + \mathbf{E} \mathbf{E}^\top \mathbf{E} \right\|_F^2 \\
&= \left\| \mathbf{Q} \bar{\mathbf{\Lambda}}^{1/2} (\bar{\mathbf{\Lambda}} - \bar{\mathbf{\Sigma}}) \mathbf{R} + \mathbf{Q} \bar{\mathbf{\Lambda}}^{1/2} \mathbf{R} \mathbf{E} \mathbf{E}^\top \right\|_F^2 + \left\| \mathbf{E} \mathbf{R}^\top \bar{\mathbf{\Lambda}} \mathbf{R} + \mathbf{E} \mathbf{E}^\top \mathbf{E} \right\|_F^2
\end{aligned} \tag{66}$$

since  $\mathbf{Q}$  is orthogonal to  $\mathbf{E}$ . The remaining part is to show at least one of the two terms is large for any  $\mathbf{W} \in \mathcal{R}'_3$  by considering the following two cases.

Case I:  $\|\mathbf{E}\|_F^2 \geq \frac{4}{25} \sigma_r(\mathbf{X}^*)$ . As  $\mathbf{E}$  is large, we bound the second term in (66):

$$\begin{aligned}
\left\| \mathbf{E} \mathbf{R}^\top \bar{\mathbf{\Lambda}} \mathbf{R} + \mathbf{E} \mathbf{E}^\top \mathbf{E} \right\|_F^2 &\geq \sigma_r^2 \left( \mathbf{R}^\top \bar{\mathbf{\Lambda}} \mathbf{R} + \mathbf{E}^\top \mathbf{E} \right) \|\mathbf{E}\|_F^2 = \sigma_r^4(\mathbf{W}) \|\mathbf{E}\|_F^2 \\
&\geq \left( \frac{1}{2} \right)^2 \frac{4}{25} \sigma_r^3(\mathbf{X}^*) = \frac{1}{25} \sigma_r^3(\mathbf{X}^*),
\end{aligned} \tag{67}$$

where the first inequality follows from Corollary 1, the first equality follows from the fact  $\mathbf{W}^\top \mathbf{W} = \mathbf{R}^\top \bar{\mathbf{\Lambda}} \mathbf{R} + \mathbf{E}^\top \mathbf{E}$ , and the last inequality holds because by assumption that  $\sigma_r^2(\mathbf{W}) \geq \frac{1}{2} \sigma_r(\mathbf{X}^*)$  and  $\|\mathbf{E}\|_F^2 \geq \frac{4}{25} \sigma_r(\mathbf{X}^*)$ .

Case II:  $\|\mathbf{E}\|_F^2 \leq \frac{4}{25} \sigma_r(\mathbf{X}^*)$ . In this case, we start by bounding the diagonal entries in  $\bar{\mathbf{\Lambda}}$ . First, utilizing Weyl's inequality for perturbation of singular values [22, Theorem 3.3.16] gives

$$\left| \sigma_r(\mathbf{W}) - \min_i \bar{\lambda}_i^{1/2} \right| \leq \|\mathbf{E}\|_2,$$

which implies

$$\min_i \bar{\lambda}_i^{1/2} \geq \sigma_r(\mathbf{W}) - \|\mathbf{E}\|_2 \geq \sqrt{\frac{1}{2} \sigma_r^{1/2}(\mathbf{X}^*)} - \frac{2}{5} \sigma_r^{1/2}(\mathbf{X}^*), \tag{68}$$

where we utilize  $\|\mathbf{E}\|_2 \leq \|\mathbf{E}\|_F \leq \frac{2}{5} \sigma_r^{1/2}(\mathbf{X}^*)$ . On the other hand,

$$\text{dist}(\mathbf{W}, \mathbf{W}^*) \leq \left\| \mathbf{Q} (\bar{\mathbf{\Lambda}}^{1/2} - \bar{\mathbf{\Sigma}}^{1/2}) \mathbf{R} + \mathbf{E} \right\|_F \leq \left\| \mathbf{Q} (\bar{\mathbf{\Lambda}}^{1/2} - \bar{\mathbf{\Sigma}}^{1/2}) \mathbf{R} \right\|_F + \|\mathbf{E}\|_F,$$

which together with the assumption that  $\text{dist}(\mathbf{W}, \mathbf{W}^*) \geq \sigma_r^{1/2}(\mathbf{X}^*)$  gives

$$\left\| \bar{\mathbf{\Lambda}}^{1/2} - \bar{\mathbf{\Sigma}}^{1/2} \right\|_F \geq \sigma_r^{1/2}(\mathbf{X}^*) - \frac{2}{5} \sigma_r^{1/2}(\mathbf{X}^*) = \frac{3}{5} \sigma_r^{1/2}(\mathbf{X}^*).$$

We now bound the first term in (66):

$$\begin{aligned} \left\| \mathbf{Q} \bar{\mathbf{\Lambda}}^{1/2} (\bar{\mathbf{\Lambda}} - \bar{\mathbf{\Sigma}}) \mathbf{R} + \mathbf{Q} \bar{\mathbf{\Lambda}}^{1/2} \mathbf{R} \mathbf{E} \mathbf{E}^T \right\|_F &\geq \min_i \bar{\lambda}_i^{1/2} \left\| (\bar{\mathbf{\Lambda}} - \bar{\mathbf{\Sigma}}) \mathbf{R} + \mathbf{R} \mathbf{E} \mathbf{E}^T \right\|_F \\ &\geq \min_i \bar{\lambda}_i^{1/2} \left( \left\| (\bar{\mathbf{\Lambda}} - \bar{\mathbf{\Sigma}}) \mathbf{R} \right\|_F - \left\| \mathbf{R} \mathbf{E} \mathbf{E}^T \right\|_F \right) \\ &\geq \left( \sqrt{\frac{1}{2}} - \frac{2}{5} \right) \left( \left( \sqrt{2} + \sqrt{\frac{1}{2}} - \frac{2}{5} \right) \frac{3}{5} - \frac{4}{25} \right) \sigma_r^{3/2}(\mathbf{X}^*) \end{aligned} \quad (69)$$

where the third line holds because  $\left\| \mathbf{E} \mathbf{E}^T \right\|_F \leq \left\| \mathbf{E} \right\|_F^2 \leq \frac{4}{25} \sigma_r(\mathbf{X}^*)$ ,  $\min_i \bar{\lambda}_i^{1/2} \geq \left( \sqrt{\frac{1}{2}} - \frac{2}{5} \right) \sigma_r^{1/2}(\mathbf{X}^*)$  by (68), and

$$\begin{aligned} \left\| \bar{\mathbf{\Lambda}} - \bar{\mathbf{\Sigma}} \right\|_F &= \sqrt{\sum_{i=1}^r (\bar{\sigma}_i - \bar{\lambda}_i)^2} = \sqrt{\sum_{i=1}^r (\bar{\sigma}_i^{1/2} - \bar{\lambda}_i^{1/2})^2 (\bar{\sigma}_i^{1/2} + \bar{\lambda}_i^{1/2})^2} \\ &\geq \left( \bar{\sigma}_r^{1/2} + \min_i \bar{\lambda}_i^{1/2} \right) \sqrt{\sum_{i=1}^r (\bar{\sigma}_i^{1/2} - \bar{\lambda}_i^{1/2})^2} = \left( \bar{\sigma}_r^{1/2} + \min_i \bar{\lambda}_i^{1/2} \right) \left\| \bar{\mathbf{\Lambda}}^{1/2} - \bar{\mathbf{\Sigma}}^{1/2} \right\|_F \\ &\geq \left( \sqrt{2} + \sqrt{\frac{1}{2}} - \frac{2}{5} \right) \frac{3}{5} \sigma_r(\mathbf{X}^*). \end{aligned}$$

Combining (65) with (66), (67) and (69) gives

$$\left\| \nabla g(\mathbf{W}) \right\|_F \geq \frac{1}{10} \sigma_r^{3/2}(\mathbf{X}^*).$$

This completes the proof of (23).

### I.3.2 Large gradient for the region $\mathcal{R}_3''$ :

By (65), we have

$$\left\| \nabla g(\mathbf{W}) \right\|_F \geq \frac{1}{2} \left\| \left( \mathbf{W} \mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T} \right) \mathbf{W} \right\|_F^2.$$

Now (24) follows directly from the fact  $\left\| \mathbf{W} \right\| > \frac{20}{19} \left\| \mathbf{W}^* \right\|$  and the following result.

*Lemma 9.* For any  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times r}$  with  $\left\| \mathbf{A} \right\| \geq \alpha \left\| \mathbf{B} \right\|$  and  $\alpha > 1$ , we have

*Proof.* Let  $\mathbf{A} = \mathbf{\Phi}_1 \mathbf{\Lambda}_1 \mathbf{R}_1^T$  and  $\mathbf{B} = \mathbf{\Phi}_2 \mathbf{\Lambda}_2 \mathbf{R}_2^T$  be the SVDs of  $\mathbf{A}$  and  $\mathbf{B}$ , respectively. Then

$$\begin{aligned} \left\| \left( \mathbf{A} \mathbf{A}^T - \mathbf{B} \mathbf{B}^T \right) \mathbf{A} \right\|_F &= \left\| \mathbf{\Phi}_1 \mathbf{\Lambda}_1^3 - \mathbf{\Phi}_2 \mathbf{\Lambda}_2^2 \mathbf{\Phi}_2^T \mathbf{\Phi}_1 \mathbf{\Lambda}_1 \right\|_F \\ &\geq \left\| \mathbf{\Lambda}_1^3 - \mathbf{\Phi}_1^T \mathbf{\Phi}_2 \mathbf{\Lambda}_2^2 \mathbf{\Phi}_2^T \mathbf{\Phi}_1 \mathbf{\Lambda}_1 \right\|_F \\ &\geq \left\| \mathbf{\Lambda}_1^3 - \mathbf{\Lambda}_2^2 \mathbf{\Lambda}_1 \right\|_F \\ &\geq \left( 1 - \frac{1}{\alpha^2} \right) \left\| \mathbf{A} \right\|^3. \end{aligned}$$

□

### I.3.3 Large gradient for the region $\mathcal{R}_3'''$ :

By (52), we have

$$\begin{aligned}
\langle \nabla g(\mathbf{W}), \mathbf{W} \rangle &= \left\langle \frac{1}{2} (\mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*\top}) \mathbf{W} + \frac{1}{2} \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*\top} \mathbf{W}, \mathbf{W} \right\rangle \\
&\geq \frac{1}{2} \left\langle (\mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*\top}) \mathbf{W}, \mathbf{W} \right\rangle \\
&\geq \frac{1}{2} \left( \|\mathbf{W}\mathbf{W}^T\|_F^2 - \|\mathbf{W}\mathbf{W}^T\|_F \|\mathbf{W}^* \mathbf{W}^{*\top}\|_F \right) \\
&> \frac{1}{20} \|\mathbf{W}\mathbf{W}^T\|_F^2
\end{aligned} \tag{70}$$

where the last line holds because  $\|\mathbf{W}^* \mathbf{W}^{*\top}\|_F < \frac{9}{10} \|\mathbf{W}\mathbf{W}^T\|_F$ . On the other hand, we have

$$\langle \nabla g(\mathbf{W}), \mathbf{W} \rangle \leq \|\nabla g(\mathbf{W})\|_F \|\mathbf{W}\| \leq \|\nabla g(\mathbf{W})\|_F \left( \|\mathbf{W}\mathbf{W}^T\|_F \right)^{1/2},$$

which further indicates that

$$\|\nabla g(\mathbf{W})\|_F > \frac{1}{20} \|\mathbf{W}\mathbf{W}^T\|_F^{3/2}.$$

This completes the proof of (25).

## J Proof of Theorem 5 (robust strict saddle for $G(\mathbf{W})$ )

We first provide several useful results regarding the deviations of the gradient and Hessian. We start with a useful characterization of RIP.

*Lemma 10.* [8] Suppose the map  $\mathcal{A} : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^p$  satisfies the  $2r$ -RIP with constant  $\delta_{2r}$ . Then we have

$$|\langle \mathcal{A}(\mathbf{C}), \mathcal{A}(\mathbf{D}) \rangle - \langle \mathbf{C}, \mathbf{AD} \rangle| \leq \delta_{2r} \|\mathbf{C}\|_F \|\mathbf{D}\|_F$$

for all matrices  $\mathbf{C}, \mathbf{D} \in \mathbb{R}^{n \times m}$  of rank at most  $r$ .

The following result controls the deviation of the gradient between the matrix sensing problem and the matrix factorization problem by utilizing the RIP of the map  $\mathcal{A}$ .

*Lemma 11.* Let  $\mathcal{A}$  satisfy the  $4r$ -RIP with constant  $\delta_{4r}$ . Then, we have

$$\|\nabla G(\mathbf{W}) - \nabla g(\mathbf{W})\|_F \leq \delta_{4r} \left\| \mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*\top} \right\|_F \|\mathbf{W}\|.$$

*Proof of Lemma 11.* We bound the deviation directly:

$$\begin{aligned}
\|\nabla G(\mathbf{W}) - \nabla g(\mathbf{W})\|_F &= \max_{\|\Delta\|_F=1} \langle \nabla G(\mathbf{W}) - \nabla g(\mathbf{W}), \Delta \rangle \\
&= \max_{\|\Delta\|_F=1} \left\langle \mathcal{A}(\mathbf{UV}^T - \mathbf{X}^*), \mathcal{A}(\Delta_U \mathbf{V}^T) \right\rangle - \left\langle \mathbf{UV}^T - \mathbf{X}^*, \Delta_U \mathbf{V}^T \right\rangle \\
&\quad + \left\langle \mathcal{A}(\mathbf{UV}^T - \mathbf{X}^*), \mathcal{A}(\mathbf{U} \Delta_V^T) \right\rangle - \left\langle \mathbf{UV}^T - \mathbf{X}^*, \mathbf{U} \Delta_V^T \right\rangle \\
&\leq \max_{\|\Delta\|_F=1} \left| \left\langle \mathcal{A}(\mathbf{UV}^T - \mathbf{X}^*), \mathcal{A}(\Delta_U \mathbf{V}^T) \right\rangle - \left\langle \mathbf{UV}^T - \mathbf{X}^*, \Delta_U \mathbf{V}^T \right\rangle \right| \\
&\quad + \left| \left\langle \mathcal{A}(\mathbf{UV}^T - \mathbf{X}^*), \mathcal{A}(\mathbf{U} \Delta_V^T) \right\rangle - \left\langle \mathbf{UV}^T - \mathbf{X}^*, \mathbf{U} \Delta_V^T \right\rangle \right| \\
&\leq \max_{\|\Delta\|_F=1} \delta_{4r} \left\| \mathbf{UV}^T - \mathbf{X}^* \right\|_F \left( \left\| \Delta_U \mathbf{V}^T \right\|_F + \left\| \mathbf{U} \Delta_V^T \right\|_F \right) \\
&\leq \delta_{4r} \left\| \mathbf{UV}^T - \mathbf{X}^* \right\|_F (\|\mathbf{V}\| + \|\mathbf{U}\|) \\
&\leq \delta_{4r} \left\| \mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*\top} \right\|_F \|\mathbf{W}\|,
\end{aligned}$$

where the second inequality utilizes Lemma 10. □

Similarly, the next result controls the deviation of the Hessian between the matrix sensing problem and the matrix factorization problem.

*Lemma 12.* Suppose  $\mathcal{A}$  satisfies the  $4r$ -RIP with constant  $\delta_{4r}$ . Then, for any  $\Delta = \begin{bmatrix} \Delta_U \\ \Delta_V \end{bmatrix} \in \mathbb{R}^{(n+m) \times r}$  the following holds:

$$|\nabla^2 G(\mathbf{W})[\Delta, \Delta] - \nabla^2 g(\mathbf{W})[\Delta, \Delta]| \leq 2\delta_{4r} \left\| \mathbf{U}\mathbf{V}^T - \mathbf{X}^* \right\|_F \left\| \Delta_U \Delta_V^T \right\|_F + \delta_{4r} \left\| \Delta_U \mathbf{V}^T + \mathbf{U} \Delta_V^T \right\|_F^2.$$

*Proof of Lemma 12.* First note that

$$\begin{aligned} & \nabla^2 G(\mathbf{W})[\Delta, \Delta] - \nabla^2 g(\mathbf{W})[\Delta, \Delta] \\ &= 2 \left\langle \mathcal{A}(\mathbf{U}\mathbf{V}^T - \mathbf{X}^*), \mathcal{A}(\Delta_U \Delta_V^T) \right\rangle - 2 \left\langle \mathbf{U}\mathbf{V}^T - \mathbf{X}^*, \Delta_U \Delta_V^T \right\rangle + \left\| \mathcal{A}(\Delta_U \mathbf{V}^T + \mathbf{U} \Delta_V^T) \right\|_2^2 - \left\| \Delta_U \mathbf{V}^T + \mathbf{U} \Delta_V^T \right\|_F^2. \end{aligned}$$

Now utilizing Lemma 10 and (30), we have

$$\begin{aligned} & |\nabla^2 G(\mathbf{W})[\Delta, \Delta] - \nabla^2 g(\mathbf{W})[\Delta, \Delta]| \\ & \leq 2 \left| \left\langle \mathcal{A}(\mathbf{U}\mathbf{V}^T - \mathbf{X}^*), \mathcal{A}(\Delta_U \Delta_V^T) \right\rangle - \left\langle \mathbf{U}\mathbf{V}^T - \mathbf{X}^*, \Delta_U \Delta_V^T \right\rangle \right| + \left| \left\| \mathcal{A}(\Delta_U \mathbf{V}^T + \mathbf{U} \Delta_V^T) \right\|_2^2 - \left\| \Delta_U \mathbf{V}^T + \mathbf{U} \Delta_V^T \right\|_F^2 \right| \\ & \leq 2\delta_{4r} \left\| \mathbf{U}\mathbf{V}^T - \mathbf{X}^* \right\|_F \left\| \Delta_U \Delta_V^T \right\|_F + \delta_{4r} \left\| \Delta_U \mathbf{V}^T + \mathbf{U} \Delta_V^T \right\|_F^2. \end{aligned}$$

□

We provide one more result before proceeding to prove the main theorem.

*Lemma 13.* [4, Lemma E.1] Let  $\mathbf{A}$  and  $\mathbf{B}$  be two  $n \times r$  matrices such that  $\mathbf{A}^T \mathbf{B} = \mathbf{B}^T \mathbf{A}$  is PSD. Then

$$\left\| (\mathbf{A} - \mathbf{B}) \mathbf{A}^T \right\|_F^2 \leq \frac{1}{2(\sqrt{2} - 1)} \left\| \mathbf{A} \mathbf{A}^T - \mathbf{B} \mathbf{B}^T \right\|_F^2.$$

### J.1 Local descent condition for the region $\mathcal{R}_1$

Similar to what used in Appendix I.1, we perform the change of variable  $\mathbf{W}^* \mathbf{R} \rightarrow \mathbf{W}^*$  to avoid  $\mathbf{R}$  in the following equations. With this change of variable we have instead  $\mathbf{W}^T \mathbf{W}^* = \mathbf{W}^{*T} \mathbf{W}$  is PSD.

We first control  $|\langle \nabla G(\mathbf{W}) - \nabla g(\mathbf{W}), \mathbf{W} - \mathbf{W}^* \rangle|$  as follows:

$$\begin{aligned} & |\langle \nabla G(\mathbf{W}) - \nabla g(\mathbf{W}), \mathbf{W} - \mathbf{W}^* \rangle| \\ & \leq \left| \left\langle \mathcal{A}(\mathbf{U}\mathbf{V}^T - \mathbf{X}^*), \mathcal{A}((\mathbf{U} - \mathbf{U}^*)\mathbf{V}^T) \right\rangle - \left\langle \mathbf{U}\mathbf{V}^T - \mathbf{X}^*, (\mathbf{U} - \mathbf{U}^*)\mathbf{V}^T \right\rangle \right| \\ & \quad + \left| \left\langle \mathcal{A}(\mathbf{U}\mathbf{V}^T - \mathbf{X}^*), \mathcal{A}(\mathbf{U}(\mathbf{V} - \mathbf{V}^*)^T) \right\rangle - \left\langle \mathbf{U}\mathbf{V}^T - \mathbf{X}^*, \mathbf{U}(\mathbf{V} - \mathbf{V}^*)^T \right\rangle \right| \\ & \leq \delta_{4r} \left\| \mathbf{U}\mathbf{V}^T - \mathbf{X}^* \right\|_F \left( \left\| (\mathbf{U} - \mathbf{U}^*)\mathbf{V}^T \right\|_F + \left\| \mathbf{U}(\mathbf{V} - \mathbf{V}^*)^T \right\|_F \right) \\ & \leq \delta_{4r} \left\| \mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T} \right\|_F \left\| \mathbf{W}(\mathbf{W} - \mathbf{W}^*)^T \right\|_F \\ & \leq \frac{\delta_{4r}}{2(\sqrt{2} - 1)} \left\| \mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T} \right\|_F^2 \end{aligned}$$

where the first inequality utilizes Lemma 10, and the last inequality follows from Lemma 13. The above result along with (53)-(54) gives

$$\begin{aligned} & \langle \nabla G(\mathbf{W}), \mathbf{W} - \mathbf{W}^* \rangle \\ & \geq \langle \nabla g(\mathbf{W}), \mathbf{W} - \mathbf{W}^* \rangle - |\langle \nabla G(\mathbf{W}) - \nabla g(\mathbf{W}), \mathbf{W} - \mathbf{W}^* \rangle| \\ & \geq \langle \nabla g(\mathbf{W}), \mathbf{W} - \mathbf{W}^* \rangle - \frac{\delta_{4r}}{2(\sqrt{2} - 1)} \left\| \mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T} \right\|_F^2 \\ & \geq \frac{1}{16} \sigma_r(\mathbf{X}^*) \text{dist}^2(\mathbf{W}, \mathbf{W}^*) + \frac{1}{32} \left\| \mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T} \right\|_F^2 + \frac{1}{4\|\mathbf{X}^*\|} \left\| \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*T} \mathbf{W} \right\|_F^2 \\ & \quad - \frac{\delta_{4r}}{2(\sqrt{2} - 1)} \left\| \mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T} \right\|_F^2 \\ & \geq \frac{1}{16} \sigma_r(\mathbf{X}^*) \text{dist}^2(\mathbf{W}, \mathbf{W}^*) + \frac{1}{160} \left\| \mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T} \right\|_F^2 + \frac{1}{4\|\mathbf{X}^*\|} \left\| \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*T} \mathbf{W} \right\|_F^2 \end{aligned} \tag{71}$$



where we utilize  $\delta_{4r} \leq \frac{1}{50}$ .

On the other hand, we control  $\|\nabla G(\mathbf{W})\|_F$  with Lemma 11 controlling the deviation between  $\nabla G(\mathbf{W})$  and  $\nabla g(\mathbf{W})$  as follows:

$$\begin{aligned}
\|\nabla G(\mathbf{W})\|_F^2 &= \|\nabla g(\mathbf{W}) + \nabla G(\mathbf{W}) - \nabla g(\mathbf{W})\|_F^2 \\
&\leq \frac{20}{19} \|\nabla g(\mathbf{W})\|_F^2 + 20 \|\nabla g(\mathbf{W}) - \nabla G(\mathbf{W})\|_F^2 \\
&\leq \frac{20}{19} \|\nabla g(\mathbf{W})\|_F^2 + 20\delta_{4r}^2 \|\mathbf{W}\|^2 \|\mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T}\|_F^2 \\
&= \frac{5}{19} \left\| \left( \mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T} \right) \mathbf{W} + \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*T} \mathbf{W} \right\|_F^2 + 20\delta_{4r}^2 \|\mathbf{W}\|^2 \|\mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T}\|_F^2 \quad (72) \\
&\leq \left( \frac{5}{19} \frac{100}{99} + 20\delta_{4r}^2 \right) \left\| \left( \mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T} \right) \mathbf{W} \right\|_F^2 + 25 \left\| \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*T} \mathbf{W} \right\|_F^2 \\
&\leq \left( \frac{5}{19} \frac{100}{99} + 50\delta_{4r}^2 \right) (\sqrt{2} + 1)^2 \|\mathbf{X}^*\| \left\| \mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T} \right\|_F^2 + 25 \left\| \widehat{\mathbf{W}}^* \widehat{\mathbf{W}}^{*T} \mathbf{W} \right\|_F^2,
\end{aligned}$$

where the first inequality holds since  $(a + b)^2 \leq \frac{1+\epsilon}{\epsilon} a^2 + (1 + \epsilon) b^2$  for any  $\epsilon > 0$ , and the fourth line follows from (52).

Now combining (71)-(72) and assuming  $\delta_{4r} \leq \frac{1}{50}$  gives

$$\langle \nabla G(\mathbf{W}), \mathbf{W} - \mathbf{W}^* \rangle \geq \frac{1}{16} \sigma_r(\mathbf{X}^*) \text{dist}^2(\mathbf{W}, \mathbf{W}^*) + \frac{1}{260 \|\mathbf{X}^*\|} \|\nabla G(\mathbf{W})\|_F^2.$$

This completes the proof of (32).

## J.2 Negative curvature for the region $\mathcal{R}_2$

Let  $\Delta = \mathbf{q}_k \boldsymbol{\alpha}^T$  be defined as in (60). First note that

$$\begin{aligned}
\left\| \Delta \mathbf{U} \mathbf{V}^T + \mathbf{U} \Delta \mathbf{V}^T \right\|_F^2 &\leq 2 \left\| \Delta \mathbf{U} \mathbf{V}^T \right\|_F^2 + 2 \left\| \mathbf{U} \Delta \mathbf{V}^T \right\|_F^2 \\
&\leq 2 \left\| \mathbf{W} \Delta^T \right\|_F^2 = 2\sigma_r^2(\mathbf{W}) \leq \sigma_r(\mathbf{X}^*),
\end{aligned}$$

where the last equality holds because  $\sigma_r(\mathbf{W}) \leq \sqrt{\frac{1}{2} \sigma_r^{1/2}(\mathbf{X}^*)}$ . Also utilizing the particular structure in  $\Delta$  yields

$$\left\| \Delta \mathbf{U} \Delta \mathbf{V}^T \right\|_F = \frac{1}{2} \left\| \phi_k \psi_k^T \right\|_F = \frac{1}{2}.$$

It follows from the assumption  $\frac{20}{19} \|\mathbf{W}^* \mathbf{W}^{*T}\|_F \geq \|\mathbf{W} \mathbf{W}^T\|_F$  that

$$\begin{aligned}
\left\| \mathbf{U} \mathbf{V}^T - \mathbf{X}^* \right\|_F &\leq \frac{\sqrt{2}}{2} \left\| \mathbf{W} \mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*T} \right\|_F \\
&\leq \frac{\sqrt{2}}{2} \left( \frac{20}{19} \left\| \mathbf{W}^* \mathbf{W}^{*T} \right\|_F + \left\| \mathbf{W}^* \mathbf{W}^{*T} \right\|_F \right) = \frac{39\sqrt{2}}{38} \|\mathbf{X}^*\|_F.
\end{aligned}$$

Now combining the above results with Lemma 12, we have

$$\begin{aligned}
\nabla^2 G(\mathbf{W})[\Delta, \Delta] &\leq \nabla^2 g(\mathbf{W})[\Delta, \Delta] + |\nabla^2 G(\mathbf{W})[\Delta, \Delta] - \nabla^2 g(\mathbf{W})[\Delta, \Delta]| \\
&\leq -\frac{1}{4} \sigma_r(\mathbf{X}^*) + 2\delta_{4r} \left\| \mathbf{U} \mathbf{V}^T - \mathbf{X}^* \right\|_F \left\| \Delta \mathbf{U} \Delta \mathbf{V}^T \right\|_F + \delta_{4r} \left\| \Delta \mathbf{U} \mathbf{V}^T + \mathbf{U} \Delta \mathbf{V}^T \right\|_F^2 \\
&\leq -\frac{1}{4} \sigma_r(\mathbf{X}^*) + \frac{39}{38} \sqrt{2} \delta_{4r} \|\mathbf{X}^*\|_F + \delta_{4r} \sigma_r(\mathbf{X}^*) \\
&\leq -\frac{1}{5} \sigma_r(\mathbf{X}^*),
\end{aligned}$$

where the last line holds when  $\delta_{4r} \leq \frac{\sigma_r(\mathbf{X}^*)}{50 \|\mathbf{X}^*\|_F}$ . This completes the proof of (33).

### J.3 Large gradient for the region $\mathcal{R}'_3 \cup \mathcal{R}''_3 \cup \mathcal{R}'''_3$ :

To show that  $G(\mathbf{W})$  has large gradient in these three regions, we mainly utilize Lemma 11 to guarantee that  $\nabla G(\mathbf{W})$  is close to  $\nabla g(\mathbf{W})$ .

#### J.3.1 Large gradient for the region $\mathcal{R}'_3$ :

Utilizing Lemma 11, we have

$$\begin{aligned}
\|\nabla G(\mathbf{W})\|_F &\geq \|\nabla g(\mathbf{W})\|_F - \|\nabla G(\mathbf{W}) - \nabla g(\mathbf{W})\|_F \\
&\geq \|\nabla g(\mathbf{W})\|_F - \delta_{4r} \left\| \mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*\top} \right\|_F \|\mathbf{W}\| \\
&\geq \|\nabla g(\mathbf{W})\|_F - \delta_{4r} \left( \frac{10}{9} \left\| \mathbf{W}^* \mathbf{W}^{*\top} \right\|_F + \left\| \mathbf{W}^* \mathbf{W}^{*\top} \right\|_F \right) \|\mathbf{W}\| \\
&\geq \frac{1}{10} \sigma_r^{3/2}(\mathbf{X}^*) - \delta_{4r} \frac{19}{9} 2 \|\mathbf{X}^*\|_F \frac{20}{19} \sqrt{2} \|\mathbf{X}^*\|^{1/2} \\
&\geq \frac{1}{27} \sigma_r^{3/2}(\mathbf{X}^*),
\end{aligned}$$

where the fourth line follows because  $\left\| \mathbf{W}^* \mathbf{W}^{*\top} \right\|_F = 2 \|\mathbf{X}^*\|_F$  and  $\|\mathbf{W}\| \leq \frac{20}{19} \sqrt{2} \|\mathbf{X}^*\|^{1/2}$ , and the last line holds if  $\delta_{4r} \leq \frac{1}{100} \frac{\sigma_r^{3/2}(\mathbf{X}^*)}{\|\mathbf{X}^*\|_F \|\mathbf{X}^*\|^{1/2}}$ . This completes the proof of (34).

#### J.3.2 Large gradient for the region $\mathcal{R}''_3$ :

Utilizing Lemma 11 again, we have

$$\begin{aligned}
\|\nabla G(\mathbf{W})\|_F &\geq \|\nabla g(\mathbf{W})\|_F - \delta_{4r} \left( \left\| \mathbf{W}\mathbf{W}^T \right\|_F + \left\| \mathbf{W}^* \mathbf{W}^{*\top} \right\|_F \right) \|\mathbf{W}\| \\
&\geq \frac{39}{800} \|\mathbf{W}\|^3 - \delta_{4r} \left( \frac{10}{9} \left\| \mathbf{W}^* \mathbf{W}^{*\top} \right\|_F + \left\| \mathbf{W}^* \mathbf{W}^{*\top} \right\|_F \right) \|\mathbf{W}\| \\
&\geq \frac{39}{800} \|\mathbf{W}\|^3 - \delta_{4r} \frac{19}{9} 2 \|\mathbf{X}^*\|_F \|\mathbf{W}\| \\
&\geq \frac{39}{800} \|\mathbf{W}\|^3 - \frac{19}{450} \|\mathbf{X}^*\| \|\mathbf{W}\| \\
&\geq \frac{1}{50} \|\mathbf{W}\|^3,
\end{aligned}$$

where the fourth line holds if  $\delta_{4r} \leq \frac{1}{100} \frac{\sigma_r^{3/2}(\mathbf{X}^*)}{\|\mathbf{X}^*\|_F \|\mathbf{X}^*\|^{1/2}}$  and the last follows from the fact that

$$\|\mathbf{W}\| > \frac{20}{19} \|\mathbf{W}^*\| \geq \frac{20}{19} \sqrt{2} \|\mathbf{X}^*\|^{1/2}.$$

This completes the proof of (35).

#### J.3.3 Large gradient for the region $\mathcal{R}'''_3$ :

To show (36), we first control  $|\langle \nabla G(\mathbf{W}) - \nabla g(\mathbf{W}), \mathbf{W} \rangle|$  as follows:

$$\begin{aligned}
&|\langle \nabla G(\mathbf{W}) - \nabla g(\mathbf{W}), \mathbf{W} \rangle| \\
&= 2 \left| \left\langle \mathcal{A}(\mathbf{U}\mathbf{V}^T - \mathbf{X}^*), \mathcal{A}(\mathbf{U}\mathbf{V}^T) \right\rangle - \left\langle \mathbf{U}\mathbf{V}^T - \mathbf{X}^*, \mathbf{U}\mathbf{V}^T \right\rangle \right| \\
&\leq 2\delta_{4r} \left\| \mathbf{U}\mathbf{V}^T - \mathbf{X}^* \right\|_F \left\| \mathbf{U}\mathbf{V}^T \right\|_F \\
&\leq 2\delta_{4r} \frac{19}{20} \sqrt{2} \|\mathbf{W}\mathbf{W}^T\|_F \frac{1}{2} \|\mathbf{W}\mathbf{W}^T\|_F = \frac{19}{20} \sqrt{2} \delta_{4r} \|\mathbf{W}\mathbf{W}^T\|_F^2,
\end{aligned}$$

where the last inequality holds because

$$\begin{aligned}\|\mathbf{U}\mathbf{V}^T - \mathbf{X}^*\|_F &\leq \frac{\sqrt{2}}{2} \|\mathbf{W}\mathbf{W}^T - \mathbf{W}^* \mathbf{W}^{*\top}\|_F \\ &\leq \frac{\sqrt{2}}{2} \left( \frac{9}{10} \|\mathbf{W}\mathbf{W}^T\|_F + \|\mathbf{W}\mathbf{W}^T\|_F \right) = \frac{19\sqrt{2}}{20} \|\mathbf{W}\mathbf{W}^T\|_F\end{aligned}$$

and

$$\|\mathbf{W}\mathbf{W}^T\|_F^2 = \|\mathbf{U}\mathbf{U}^T\|_F^2 + \|\mathbf{V}\mathbf{V}^T\|_F^2 + 2\|\mathbf{U}\mathbf{V}^T\|_F^2 \geq 4\|\mathbf{U}\mathbf{V}^T\|_F^2$$

by noting that

$$\|\mathbf{U}\mathbf{U}^T\|_F^2 + \|\mathbf{V}\mathbf{V}^T\|_F^2 - 2\|\mathbf{U}\mathbf{V}^T\|_F^2 = \|\mathbf{U}^T\mathbf{U} - \mathbf{V}^T\mathbf{V}\|_F^2 \geq 0.$$

Now utilizing (70) to provide a lower bound for  $\langle \nabla g(\mathbf{W}), \mathbf{W} \rangle$ , we have

$$\begin{aligned}|\langle \nabla G(\mathbf{W}), \mathbf{W} \rangle| &\geq \langle \nabla g(\mathbf{W}), \mathbf{W} \rangle - |\langle \nabla G(\mathbf{W}) - \nabla g(\mathbf{W}), \mathbf{W} \rangle| \\ &> \frac{1}{20} \|\mathbf{W}\mathbf{W}^T\|_F^2 - \frac{19}{20} \sqrt{2} \delta_{4r} \|\mathbf{W}\mathbf{W}^T\|_F^2 \\ &\geq \frac{1}{45} \|\mathbf{W}\mathbf{W}^T\|_F^2,\end{aligned}$$

where the last line holds when  $\delta_{4r} \leq \frac{1}{50}$ . Thus,

$$\|\nabla G(\mathbf{W})\|_F \geq \frac{1}{\|\mathbf{W}\|} |\langle \nabla G(\mathbf{W}), \mathbf{W} \rangle| > \frac{1}{45} \|\mathbf{W}\mathbf{W}^T\|_F^{3/2},$$

where we utilize  $\|\mathbf{W}\| \leq (\|\mathbf{W}\mathbf{W}^T\|_F)^{1/2}$ . This completes the proof of (36).

## References

- [1] S. Aaronson. The learnability of quantum states. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 463, pages 3089–3114. The Royal Society, 2007.
- [2] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning sparsely used over-complete dictionaries. In *Conference on Learning Theory (COLT)*, pages 123–137, 2014.
- [3] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi. Dropping convexity for faster semi-definite optimization. *arXiv preprint*, 2015.
- [4] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221*, 2016.
- [5] S. Burer and R. D. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.
- [6] S. Burer and R. D. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.
- [7] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.
- [8] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathématique*, 346(9-10):589–592, 2008.
- [9] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.
- [10] E. J. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.
- [11] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

- [12] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.
- [13] Y. Chen and E. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, pages 739–747, 2015.
- [14] G. S. Chirikjian and A. B. Kyatkin. *Harmonic Analysis for Engineers and Applied Scientists: Updated and Expanded Edition*. Courier Dover Publications, 2016.
- [15] A. R. Conn, N. I. Gould, and P. L. Toint. *Trust region methods*. SIAM, 2000.
- [16] M. A. Davenport and J. Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.
- [17] M. Fazel, H. Hindi, and S. Boyd. Rank minimization and applications in system theory. In *American Control Conference*, volume 4, pages 3273–3278. IEEE, 2004.
- [18] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.
- [19] R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. *arXiv preprint arXiv:1605.07272*, 2016.
- [20] Z. Harchaoui, M. Douze, M. Paulin, M. Dudik, and J. Malick. Large-scale image classification with trace-norm regularization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3386–3393. IEEE, 2012.
- [21] N. Higham and P. Papadimitriou. Matrix procrustes problems. *Rapport technique, University of Manchester*, 1995.
- [22] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012.
- [23] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, pages 665–674. ACM, 2013.
- [24] C. Jin, S. M. Kakade, and P. Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. *arXiv preprint arXiv:1605.08370*, 2016.
- [25] C. R. Johnson, K. Okubo, and R. Reams. Uniqueness of matrix square roots and an application. *Linear Algebra and Its Applications*, 323(1):51–60, 2001.
- [26] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent converges to minimizers. *University of California, Berkeley*, 1050:16, 2016.
- [27] K. Lee, N. Tian, and J. Romberg. Fast and guaranteed blind multichannel deconvolution under a bilinear system model. *arXiv preprint arXiv:1610.06469*, 2016.
- [28] Q. Li and G. Tang. The nonconvex geometry of low-rank matrix optimizations with general objective functions. *arXiv:1611.03060*, 2016.
- [29] X. Li, S. Ling, T. Strohmer, and K. Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *arXiv preprint arXiv:1606.04933*, 2016.
- [30] X. Li, Z. Wang, J. Lu, R. Arora, J. Haupt, H. Liu, and T. Zhao. Symmetry, saddle points, and global geometry of nonconvex matrix factorization. *arXiv preprint arXiv:1612.09296*, 2016.
- [31] Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2009.
- [32] I. Panageas and G. Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. *arXiv preprint arXiv:1605.00405*, 2016.
- [33] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. *arXiv preprint arXiv:1609.03240*, 2016.
- [34] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.
- [35] N. Srebro, J. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems*, pages 1329–1336, 2004.

- [36] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method. *arXiv preprint arXiv:1511.04777*, 2015.
- [37] J. Sun, Q. Qu, and J. Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.
- [38] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *arXiv preprint arXiv:1602.06664*, 2016.
- [39] S. Tu, R. Boczar, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.
- [40] L. Wang, X. Zhang, and Q. Gu. A unified computational and statistical framework for nonconvex low-rank matrix estimation. *arXiv preprint arXiv:1610.05275*, 2016.
- [41] T. Zhao, Z. Wang, and H. Liu. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567, 2015.
- [42] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin. Global optimality in low-rank matrix optimization. *arXiv preprint arXiv:1702.07945*, 2017.